Mental Tests in Clinics for Children

Grace H. Kent

136.702
K37m
c.2

PRESENTED TO
COUNSELING AND GUIDANCE CENTER
IN MEMORY OF
Justin E. Harlow, Jr.
BY HIS FAMILY

UNIVERSITY
OF FLORIDA
LIBRARIES

Digitized by the Internet Archive
in 2010 with funding from
Lyrasis Members and Sloan Foundation

http://www.archive.org/details/mentaltestsincli00kent

TEXTBOOKS ON PSYCHOLOGY

J. P. GUILFORD, *University of Southern California*
General Editor

~~◇~~

General Psychology. BY *J. P. Guilford.*

Studies in Elementary Psychology. A Manual and Workbook. BY *J. P. Guilford and William E. Walton.*

Fields of Psychology. (In association with Eminent Specialists.) 2nd Edition.

Physiological Psychology. BY *G. L. Freeman, Cornell University.*

Mental Tests in Clinics for Children. BY *Grace H. Kent, Vermont State Hospital.*

~~◇~~

PUBLISHED BY

D. VAN NOSTRAND COMPANY, INC.

MENTAL TESTS

IN CLINICS FOR CHILDREN

by

GRACE H. KENT

Vermont State Hospital
Waterbury, Vermont

D. VAN NOSTRAND COMPANY, Inc.

TORONTO NEW YORK LONDON

NEW YORK

D. Van Nostrand Company, Inc., 250 Fourth Avenue, New York 3

TORONTO

D. Van Nostrand Company (Canada), Ltd., 228 Bloor Street, Toronto

LONDON

Macmillan & Company, Ltd., St. Martin's Street, London, W.C. 2

Copyright, 1950, by D. Van Nostrand Co., Inc.

Published simultaneously in Canada by
D. Van Nostrand Company (Canada), Ltd.

All Rights Reserved

This book, or any parts thereof, may not be reproduced in any form without written permission from the author and the publishers.

PRINTED IN THE UNITED STATES OF AMERICA

BY GEORGE S. FERGUSON COMPANY
PHILADELPHIA, PA.

PREFACE

This is intended as an auxiliary textbook for advanced students in clinical psychology, written to encourage in them a critical and experimental attitude in their use of mental tests. It has been my wish to avoid any considerable overlap with the writings of others, and to present a point of view not generally recognized.

Ever since Binet gave us the idea of a "mental yardstick," it has been customary to evaluate individual achievement in terms of statistical criteria. Parallel with this trend there has been a tendency to lose sight of the individual. The norms of standard tests are based upon majority findings, and are doubtless applicable to the majority of children. It does not follow, however, that they are applicable to the exceptional child. Most of the children referred to the clinic are exceptional in some way, and we have no standard instrument that offers the exceptional child a fair opportunity to show what he can do. We examine him by a scale that may be grossly unfair to him, and report his achievement as compared with that of normal children.

The psychological test holds vast possibilities for aiding us in understanding the child and his problems, but it is used primarily for obtaining a numerical rating. I have the utmost confidence in the potential value of the test as a clinical instrument, but am less confident of the value of norms. In recent years I have come to believe that the development of norms might better be restricted to tests that admit of objective evaluation; and I should like to see all other tests used as instruments for individual observation, the findings to be reported qualitatively.

Even for the most nearly objective tests that we have, I should prefer to recognize for the report only three levels of achievement: high, average, and low. The examiner usually knows nothing about the subject beyond what he can learn in a formal and hurried one-hour interview. If he can place a child in one of three categories, this is as much as can reasonably be expected. The danger of putting misleading information on permanent record would be greatly re-

duced if the quantitative report were limited to what the examiner can state with some measure of assurance. Furthermore, a descriptive report of the responses to unstandardized tasks might be used more freely for what they would reveal of the child's personality.

It is only by way of quotation—implied if not expressed—that I refer to the "intelligence quotient." I never use the term "intelligence" as applied to anything that can be measured. My use of the term "mental-age" is a concession to convenience, and is not intended to imply acceptance of the concept. What we can measure by any scale or battery does not constitute a child's total mental capacity. The rating-by-age, however, has its own significance for what is measured by means of any test. Because of the wide recognition Binet's term has received over a forty-year period, it has become practically indispensable as a common denominator for tests having age norms. It is as such that I have adopted it for the tests standardized by me.

Of necessity, my development of norms for publication has placed me in an ambivalent position and has involved a vast amount of rationalization on my part. It is not for me to know whether the serviceability of my tests is sufficient to balance the danger that their norms will be misused. I have little interest in the "probable error," being more deeply concerned with reducing the *possible error* in the application of the norms to the individual subject. It is of course my wish that any test of mine be used with discrimination, not mechanically.

The first four chapters of this, written in the spring of 1948, were used in lectures at University of Miami, Florida. The remaining chapters were written in 1949 at Vermont State Hospital, Waterbury, Vermont. I am indebted to the two institutions for permitting me to have time for writing, especially to Dr. Rupert Chittick of the Vermont institution. For critical reading of the text, chapter by chapter, I am grateful to Ernest Kent and Robert Allen. I accept responsibility for the content of the book. So far as I am aware, no student in the field will be in full agreement with me concerning the light esteem in which I hold statistical criteria.

GRACE H. KENT.

Arlington, Vermont

CONTENTS

CHAPTER		PAGE
PREFACE		v
FOREWORD		xi
I.	WHY CHILDREN ARE REFERRED TO THE CLINIC	1
	1. Introductory Note	1
	2. Misbehavior in School	1
	3. The Mentally Retarded Child	2
	4. The Gifted Child	3
	5. Difficulties in Learning to Read	4
	6. Behavior Disorders Based Upon Organic Condition	5
	7. Problems Following Prolonged Illness	7
	8. Suitability for Adoption	9
	9. Children for Foster-Home Placement	10
	10. Misconduct in the Community	10
	11. Miscellaneous Clinical Problems	11
II.	THE CLINICAL EXAMINATION	13
	1. Importance of Making the Examination Attractive	13
	2. Distribution of Unpleasant Tasks	15
	3. The Place of Praise in the Examination	16
	4. Pencil-and-Paper Tests	17
	5. The Preschool Child	18
	6. What to Observe in the Subject's Performance	19
	7. Physical Equipment of the Examining Office	20
III.	SOURCES OF ERROR IN USE OF TESTS	23
	1. Statistical Fallacy	23
	2. Constant Error in Clinical Use of Norms	24
	3. Interference of Emotional Factors	25
	4. Erroneous Estimate of the Subject's Cooperation	26
	5. Fatigue of the Examiner	27
	6. The Subjective Factor in Test Presentation	28
	7. Estimate of "Intelligence Quotient"	29
	8. Other Possible Errors	30
	9. Practical Effects of the Various Inaccuracies	31

CHAPTER PAGE

IV. The Composite Scale Versus the Battery 34
1. Explanation of Terms 34
2. Weakness of the Composite System, in the Clinic 35
3. Strength and Weakness of the Battery 37
4. Structure of Performance Scales 39

V. Language and Performance Tests, With and Without Timing 45
1. Distinction Between Verbal and Nonverbal Tests 45
2. Significance of Verbal and Nonverbal Tests .. 46
3. Oral and Written Tests 50
4. Timed and Untimed Tests 51
5. Distinctive Features of Performance Tests .. 54
6. Significance of Speed in Performance Tests .. 56
7. Diagrammatic Classification of Tests 59

VI. Use of Unstandardized Tests 61
1. "Shock Absorber" and "Buffer" Tasks 61
2. Advantages of the Nonstandard Task, for Informal Use 62
3. Effects of Standardization Upon Test Development 64
4. Misuse of Standard Tests 67
5. Subjective and Objective Criteria 68
6. Practical Studies in Group Classification 70
7. The Test as a Personal Instrument 71

VII. The Psychological Report 74
1. General Plan 74
2. The Introductory Section 76
3. Findings by Language Tests 78
4. Findings by Performance Tests 81
5. Other Observations and Impressions 84

VIII. Criteria for Construction of Clinical Tests ... 86
1. Long-Range Plan for a Flexible Scale 86
2. Set-up for the Orally Presented Sub-Test ... 89
3. Criteria for Items of an Oral Scale 92
4. Criteria for Evaluation and Weighting 94

CHAPTER PAGE

5. Criteria for Arithmetical Scales 96
6. The Vocabulary Test 99
7. Criteria for Written Tests 102
8. Criteria for Nonverbal Scales 105
9. Criteria for Standardization 106
10. Validation of a Battery 110
11. Possibilities of the Formboard 113
12. Development of Color-Cube Scales 115
13. General Criteria for Performance Tests 118

IX. Serviceable Scales and Sub-Tests 121
Introductory Note 121
1. The Detroit Tests 121
2. The Kuhlmann Tests 127
3. The "Emergency" Tests, Oral and Written .. 128
4. Vocabulary Tests 131
5. Language Tests for Adolescent Levels 132
6. Nonverbal Scales 134
7. Formboard Tests 138
8. The Color-Cube Test 146
9. Preschool Performance Tests 147
10. The Young Mazes 148

X. Home Manufacture of Materials 150
1. Language Tests 150
2. Performance Tests 152

Appendix A 161

Appendix B 162

Bibliography 165

Index 175

FOREWORD

In this volume Dr. Kent has recorded in very clear and succinct form her conclusions from a lifetime of fruitful experiences in psychological clinics. It is quite important that the suggestive value of those experiences be preserved for clinical posterity. It is for the very reason that her conclusions do not always agree with orthodox teachings and practices in test usage that what she has to say should be made generally available. Whether it is on the subject of test construction, test standardization, or test administration, she has sound advice to offer the research psychologist, the psychometrician, and the responsible clinical psychologist.

The reader will find that Dr. Kent is very critical of the use of the IQ, on general grounds, and, more particularly, its misuse by laymen and even by psychologists. She much prefers the standard of a mental-age scale, but in some places expresses preferences for centile rank based upon age norms. In this connection, and in others, when statistical issues are involved, her comments are restricted to superficial levels. The issues run much deeper and cannot be solved at those levels.

Her discontent with the IQ goes further. She is definitely against the use of any single composite score to represent degree of intellect, preferring a battery of tests that provide several different scores, each of which can be referred to standards. Not only does this offer more information about individuals but also greater flexibility in test practice.

In her recognition of many different aptitudes rather than a single intellect, Dr. Kent demonstrates the need for factor theory and factor methods without explicitly saying so. Keeping her eyes on the goal of better adjustment of the child, educationally, socially, and vocationally, she finds it sufficient to make the broader distinctions of verbal vs. nonverbal, speed vs. power, and written vs. oral, with casual references to traditional functions such as memory, perception, and reasoning. Following this trend to its logical conclusion, one would find in factorial thinking and practices more rigorous answers to many of the testing problems.

The author emphasizes, always, that tests should be adapted to the child; with his interests, his handicaps, and his attitude toward the testing situation being taken into account. This places a greater demand upon the number and kinds of tests that must be ready for use in the clinician's repertoire. Even unstandardized tests should be used, according to the author, not only at times but frequently. These have motivating functions, provide the basis for qualitative observations, and keep the examiner's observational habits alive. Not all test results should be expressed in terms of numbers; in fact, too much dependence upon objective and routinely obtained information is regarded as often unfortunate. The psychometrician should be a psychologist; not merely a laboratory technician. He should not be a slave to standardization where standards do not apply. One might sum this up by saying that testing is an art, and, although often based upon science, it can never dispense with the artist.

With most of these recommendations only the extreme positivist would completely disagree. In clinical practice, as in medical practice, there will always be points at which science ends and art begins. It is well for both scientist and clinician to recognize this general principle and for the latter to know where those points are. It is hoped that, in addition to its practical value to the operating clinical examiner, this volume will help to further that recognition and that knowledge.

J. P. Guilford.

CHAPTER I

WHY CHILDREN ARE REFERRED TO THE CLINIC

1. Introductory Note

A large proportion of children referred for clinical study are sent by school authorities. Many others are sent by the various social agencies or by practicing physicians. Some are referred by unofficial social centers, such as a church. Child-placing agencies require a psychological examination as a routine measure before placing a child in a foster home. Juvenile offenders are sent by the court, and a few children are brought by parents or guardians.

There is some overlap between the work of the psychologist and that of the social worker. In ascertaining the facts about a child's past and in explaining the recommendations for his guidance in the future, sometimes the psychologist has direct contact with the parents and sometimes he depends upon the social worker. In any case, the psychological tests have a purpose beyond finding the child's mental level. The understanding of a child's capacities is but a means to an end. Even in a manual on the use of tests, there is a place for the broader and more constructive aspects of clinical studies.

2. Misbehavior in School

The school frequently refers to the clinic a child who does not adapt himself to school routine and who perhaps keeps the whole class in an uproar of hilarity. The chief causes of disorderly behavior are very different and in different children may be diametrically opposed. The mentally retarded child is not kept properly occupied, because most of the instruction goes over his head. On the other hand, the child who is mentally ahead of his class has too much time for play because the work is too easy for him. The child who is mentally ill is irresponsible, and his behavior is unpredictable.

These categories, however, do not cover all cases of troublesome behavior at school. A healthy child is normally more active than he is expected to be in the schoolroom. Some violations of schoolroom propriety are accepted by most teachers as inevitable, and it is the exceptional case that is referred for clinical study. The chief purpose of the psychometric examination is to ascertain how much can reasonably be expected from the child.

3. The Mentally Retarded Child

A child's failure to make normal progress in school is frequently the first recognized indication of a mild mental retardation. Very rarely are the parents willing to face the fact that a child is of subnormal capacity. It may be clearly apparent to an unbiased observer that a child is very far from normal, but the mother can offer endless explanations for the apparent retardation and is sure that it is merely a question of finding the right school or tutor. "If Helen Keller could be taught the use of language, surely a child who has his sight and hearing can learn. I want to find an Anne Sullivan for my child."

There are all degrees of mental retardation, from the custodial case to the child who with proper training can become self-supporting. It is for children of the upper level of retardation that the psychological examination is most helpful. The mental test may serve to point out in what respects the child is strongest and what type of training will be most constructive. Every aptitude that he possesses should be developed. Almost any child above the level of requiring institutional care can be taught to use his hands, even if he cannot acquire a useful knowledge of reading. There is a wide field for nonacademic educational endeavor, and there is no discount on the mental development that can be achieved by means of household tasks. The home furnishes the opportunity for the child's training, if the mother can be patient. The retarded child may learn very slowly, and the instructions may have to be repeated many times before he can perform the task without supervision, but it is worth while in the long run. The child may find his greatest happiness in

being helpful and in feeling that his contribution to the family economy is recognized and appreciated.

Instead of trying to force a child into a vocation offering no possibility of success, the parents should help him to find satisfaction in doing something that he can do well. The attitude of the parents may determine whether the child is to go through life feeling inferior to his brother or whether he can take pride in doing excellent work in a field suited to his capacity. The greatest service which the clinic can perform for the retarded child is to help the parents to scale down their ambition for him and encourage him to do something within his reach.

4. The Gifted Child

Children of capacity above the average present problems of different types.

Adjustment in school is difficult for the superior child because there is no group in which he will be among his equals both intellectually and socially. It is only when he is large for his age that he stands a good chance of being accepted socially by the older children. Even with this advantage he usually lacks the emotional maturity of his older classmates. If kept in the grade in which he belongs by age, he may lose interest in school work because it is too easy to keep him busy. He may become a disciplinary problem, not from any malicious wish to be annoying but merely for lack of occupation. Finding that he can make a creditable showing in school without doing any serious work, he may fall into the habit of idling through high school, with full expectation of going in the same way through college and through life.

On the playground he may be ambitious for leadership, developing a domineering attitude which is resented by his companions. If rejected by the group, he may withdraw from all social activities and spend his free time reading or experimenting with his chemical set. Although his leisure activities may be educationally constructive, his parents note with anxiety that he is becoming solitary in his tastes and habits. It is a child of this type, especially an "only" child, who is likely to be brought to the clinic by his parents rather

than by the school. It is occasionally observed that the child talks very glibly about his "inferiority complex."

The question whether a child of superior ability should be advanced in school beyond his age is usually settled while the child is too young to have any voice in the decision. In making the choice for him, the parents and teachers should give due weight to his physical development. A large child can make adjustment to a group of older children much more easily than can an undersized one. It may safely be recommended that a well-developed child under ten be advanced in school, but care must be taken to have him master any essential item in the omitted grade, such as the multiplication table.

When a double promotion is contraindicated, the child may be assigned extra work along the lines of his own interests. With the aid of the school library and with the cooperation of his parents, it should be possible to find suitable occupation for him during study hours.

5. Difficulties in Learning to Read

Aptitude for acquiring reading knowledge is by no means identical with mental capacity in general, however high the correlation between them. A child of very low mental level can occasionally be taught to read, just as a parrot can be taught to speak. Also, there are children of near-normal mental level who have been passed up to the sixth grade—usually after spending two years in each of the early grades—without having learned to read more than a few detached words.

Nonreading children are referred for clinical study in considerable numbers, and in each case it is important to learn as much as possible about the child's mentality. In all verbal tests the child who cannot read is at a disadvantage as compared with the reading child of equal native ability. Most children begin in third or fourth grade to acquire some knowledge through reading; consequently the norms of language tests are based upon the achievement of children who can read. The nonreading child of reading age cannot be expected to achieve a normal rating for his age in verbal tests, but the

performance tests may afford him a fair opportunity to show what he can do.

For mentally retarded children, reading knowledge is of minor importance. Those who acquire it easily should by all means have the opportunity, if only for their own entertainment and for the satisfaction it will give their parents. But the retarded child who does not learn easily may better spend his time learning something that is within his grasp. Reading knowledge is for him a luxury, not to be forced on him at too great a cost.

The near-normal child who does not learn to read in the elementary grades is in a different category. Illiteracy is for him almost as serious a handicap as mental deficiency itself. Not only will his occupational and social place in the community be restricted because of his inability to read, but also his personality may be so crippled that he will never develop a normal sense of self-respect. No effort should be spared to help him acquire a serviceable reading knowledge.

The findings of the mental test are of inestimable value in determining whether the nonreader should be given special instruction in remedial reading, or whether his education should be primarily along manual lines. The proper time to refer the child for clinical study is during or immediately following his first year in school. Any child who spends a year in first grade without making any progress in reading should be given an exhaustive examination to find the cause of his failure to read. If mental retardation can be excluded, a year of special coaching in reading may enable him to enter the second grade the following year and subsequently to make normal progress.

6. Behavior Disorders Based Upon Organic Condition

No more desperate problem reaches the clinic than that of the child whose brain has been seriously damaged by birth trauma or by some specific infection such as encephalitis lethargica. With or without the mental retardation which may result from the brain condition, the child may be so erratic and irresponsible in his behavior as to strain the resources of school and home almost beyond endur-

ance. Not infrequently the child is excluded from school as being completely unmanageable and also as being incapable of learning anything. The referral to the clinic may come from a parent who claims the right to place in school a child who is more like a wild animal than a human being; or it may come from the school, on the ground that the teacher is ready to resign unless the child is withdrawn from school.

The most characteristic symptom is a lack of capacity for inhibition. The child acts at all times on the impulse of the moment, frequently scratching and biting anyone who offends him. Closely related to the lack of inhibition is the incapacity for attention. Neither the teacher nor the psychological examiner can hold the child's attention long enough to impart instruction or even to find out what instruction he is capable of receiving.

Thus the examination presents great difficulty, and the findings can be evaluated only with reservations. But, although the results may throw little light on the child's mental level, the psychological examination may be of value in indicating the starting point and direction of the training that may be attempted. With this in view it is well to let the mother witness the examination.

The techniques of the occupational therapist are employed (71). Play-tasks are presented rather casually, one after another in quick succession, to find something that appeals to the child strongly enough to elicit constructive effort. The task may be as simple as stringing the kindergarten beads, but it is a real achievement to get him to string them instead of scattering them on the floor. After he has followed one suggestion, his resistance will be appreciably weakened. Instead of picking up the beads indiscriminately, he may then be induced to string them according to a specified pattern. Gradually, with occasional rest periods for free play, the examiner may present increasingly difficult tasks. The child may thus learn that constructive activity is more interesting than random or destructive expenditure of energy. The mother may be amazed to see how much can be accomplished by gentle persuasion.

The number of birth trauma cases referred for clinical study is very large. In many instances these are the children who would not have survived if they had been born a generation earlier.

The manifestations of either head injury or specific infection are as varied as possible. A child may be reduced to idiocy, or he may make a full recovery. The capacity for responding to training differs widely, but the training must in any case be highly specialized. Sick children do not necessarily respond to the same disciplinary measures that influence the conduct of a normal child. The parents, if they are fully aware of the severity of a child's condition, may be almost frantic with grief and anxiety. Inevitably, their emotional state renders them unfit to preserve the poise that is essential in training a sick child. Sometimes, again, they may fail to recognize that the child is ill and may try to hold him responsible for his conduct. Worst of all, the two parents may disagree in their methods of dealing with the child. Consistency of treatment, important enough in the training of a healthy child, is even more urgent in the treatment of a child whose illness makes him unduly erratic in his behavior.

In general, children who are mentally ill respond better to institutional training than to any effort at training on the part of the parents. There are only a few mental hospitals for children, however, and a psychotic child is not usually considered eligible for admission to the state training school for defectives. The majority of these children are cared for at home in their early years. If they become too troublesome to be kept in the community, they may have to be placed in the state hospital with adult patients. Parents will consider this only as a last resort.

When a child is referred by the school with reference to his behavior rather than his school achievement, an organic disorder is immediately suspected. Inquiries relating to his birth or to illness in early childhood frequently bring information that strengthens the suspicion of organic disorder even when there is no final proof.

7. Problems Following Prolonged Illness

The child who has suffered a severe illness is very likely to present a problem after recovery. The nursing care which he has received during the acute stage of the illness has been accepted by him as something that belongs to him by natural right. His craving for

special attention and indulgence has become habitual, and he has developed an emotional dependence upon his parents which is all out of proportion to his physical needs after recovery. In brief, he is a "spoiled" child. No one can say just when his demands upon the family ceased to be reasonable. It is after he has become a family tyrant that the aid of the clinic is sought.

The clinic can do little or nothing directly for the child, except to ascertain his mental level for its bearing upon the extent to which he should be held responsible for his conduct. Beyond this, it is the mother who can profit by educational measures. The mother must be led to see that she is doing her child a wrong in permitting him to employ infantile techniques for making himself the center of attention. She must be cautioned not to ask the child how he is feeling, nor to remind him in any way that he has been sick. In order that she may be helped to develop a more healthy attitude on her own part, she should be discouraged from talking about the child's illness to her friends, even when the child is not present. Her own attention should be centered upon her thankfulness that he has made so good a recovery, rather than upon what she suffered during his sickness.

In case of chronic illness, the child need not be permitted to grow up believing that he is exempt from all moral responsibility. Even a sick child has his duties and obligations. It is as important to his own happiness that he learn to bear pain uncomplainingly and that he form the habit of showing proper consideration for others, as that he be taught to read. The child's whole future depends very largely upon the courage with which the mother meets the situation.

A mother whom the writer knew in 1910 has for many years served as a model to parents of sick children. Her only child, a boy of nine years, had lost one eye and his right arm in an accident that all but took his life. Having no desire for life, the child gave his physicians little cooperation in their efforts to save him. When he was brought home from the hospital he seated himself by the window and said: "There's nothing I can do now, except look out the window." The mother did not trust herself to reply. She sought the privacy of her room, not to weep but to make a solemn resolution: "My child may go through life with a maimed body, but his spirit *shall not* be crippled."

From that moment she made it her first business to practice one-hand activities of all kinds, to see what she could learn to do with her own left hand. While the boy was at school she went far out of her way to leave undone any household task that could be performed with one hand, so that he might always find some little job waiting for him when he came home. Each day she had something new for him. She presented tasks of gradually increasing difficulty, requesting his help as casually as if she had entirely forgotten about the accident. When he protested that he "couldn't do it," she laughed and assured him that anyone could do it with one hand. She told him that the use of two hands for a one-hand job was merely a matter of habit, and that it was up to him to get into the habit of doing things with one hand. Before long it became for him a sporting proposition to see how many things he could do with his left hand.

The writer last saw this boy about a year after the accident, when visiting the family at their summer camp. The skill he had developed was truly amazing. He was an excellent swimmer, and was proficient in many competitive outdoor sports. It was with admiration, not with pity, that one watched him at play with the other boys. He was the most active one of the group, and by far the best coordinated. He used his one hand so well that the other hand was scarcely missed.

This mother denied herself the personal satisfaction of giving her child any nursing care or service beyond his actual needs, and thus built up in him a spirit of self-reliance. An over-sympathetic mother might easily have developed in him an attitude of invalidism.

8. Suitability for Adoption

Very young children who are available for adoption are occasionally referred to the clinic for such prognosis as can be offered concerning their capacity for mental development. The younger the child, the less definite the information and the greater the risk to be assumed by the adoptive parents. Also, the younger the child, the more truly will it come to be their very own. It is for the parents to decide whether they are willing to accept a heavier risk for the sake of a greater reward, but they are entitled to be fully informed concerning the nature of the risk. The psychologist should be very cautious about predicting the future of a child who is examined while under a year old. The parents can more safely rely upon the advice of a physician.

For a child of about two years it is possible—although by no means certain—that the psychological examination will have real significance as indicating that he is a good risk for adoption. It is of utmost urgency that the two-year-old be examined before being placed on trial for adoption, not after having spent a few weeks in the home of the adoptive parents. In a child's development there is no more fascinating stage than that period in which he is learning to talk, picking up new words each day. The parents who want a child are all set to become attached to any child that is offered them. Let him stay with them a month, and they will no more give him up voluntarily than if he were their own child. For better or worse, he has become a member of the family. Social agencies responsible for placement of foundlings should keep this in mind and should refer the children for examination before placing them for adoption.

9. Children for Foster-Home Placement

One of the most common social situations met in clinical studies of children is the broken home. The mother who works to support her children is not usually able to make a home for them at the same time. Except when they can be cared for by relatives, it may be necessary to find a boarding home for them. Foster homes under the supervision of social agencies are usually to be preferred to any home that a mother could find unaided.

Child Welfare agencies refer the children for exhaustive clinical study as an aid to placing each child in the home best suited to his individual needs. Usually it is desirable to find a home that will accommodate all the children of a family, but there are exceptional cases in which it is highly desirable to separate the siblings. When the older child is retarded and the younger one advanced in school, it may be recommended that they be in different schools.

10. Misconduct in the Community

The juvenile or adolescent offender is very likely to be referred for clinical study before his case is brought up in court. The extent

to which the child may be held responsible for his conduct depends largely upon his mental level as determined by tests, but this is only one of many factors to be considered. A study of the home situation as furnished by the social worker throws more light, both on the causes that are leading this child into a life of crime and on the course of treatment that may save him for useful citizenship, than do the results of tests. None the less, the tests do reveal something of his capacity for responding to favorable influences. A study of his aptitudes and interests as shown by tests may indicate the foundation upon which a program of rehabilitation should be built.

11. Miscellaneous Clinical Problems

Additional reasons for referring children to the clinic—either less urgent or less frequent or less constructive—may be grouped together.

Defective speech is a frequent accompaniment of mental retardation, but it occurs occasionally in a child not mentally retarded. The speech case is usually referred for treatment to a special teacher, who may learn from the general psychological examination how much cooperation the child is capable of giving in correcting his speech. The mother can be helpful in teaching a very young child to speak correctly, but a child of school age usually responds better to someone outside the family.

When a child is referred for emotional instability, it is usually found that the mother is emotionally immature. She can best help the child by building up her own capacity for self-control. An adolescent girl who cries too easily may be advised in the presence of her mother to count up to fifty when she feels upset, or to repeat a stanza of a favorite poem or hymn. There is a good chance that the mother will accept the suggestion for herself and that the two will work together in developing techniques for acquiring better poise.

Some of the other problems for which mothers bring their children to the clinic are sibling jealousy, enuresis, refusal of food, tantrums, nail biting, finger sucking, or masturbation. In nearly all such cases the clinic's work is to be done with the mother rather than

directly with the child. Frequently it seems to the examiner that the mother's obsessive anxiety is serving only to keep the child's attention centered upon something which he might better be allowed to forget. No promise can be made, however, that the undesirable habit will take care of itself. The mother has already found that it is not overcome by nagging, and therefore she may be ready to try something else.

CHAPTER II

THE CLINICAL EXAMINATION

1. Importance of Making the Examination Attractive

The first essential of a clinical examination is that the interview be a comfortable one for the subject. He may be a timid child who clings fearfully to his mother's hand, and whose attention can best be caught by the presentation of a toy; or he may be a self-conscious adolescent who has been referred by the court and who keenly resents the whole procedure. In any case, the examiner's first task is to put the subject at his ease.

The child should be allowed a little time in which to adapt himself to the situation. The first question should be a casual one concerning some play interest, and the first performance task should be something that does not carry a score. All tasks that constitute a part of the formal examination should be held in reserve until the subject feels at home with the examiner and is ready to cooperate.

In a large majority of juvenile cases, the unwilling subject responds more readily to a task to be performed with his hands than to a question calling for a verbal answer. Except when a child has some interest of his own that he is eager to talk about, speech comes rather hard and is highly conducive to self-consciousness. A performance task, on the other hand, gives the child an excuse for the finger activity which would otherwise be used for restless fidgeting. His interest in the task takes his attention from himself as a subject-under-examination. The examiner should have at hand a varied assortment of play-tasks, ranging in difficulty from the nursery-school peg board to the mechanical problem requiring some real effort on the part of the high-school boy. Thus equipped, one can open the examination by presenting a formboard with the casual request: "Let me see how quickly you can get these blocks into their places."

The ideal introductory task, for a given subject, is one that will be performed successfully but not too easily. The task that proves to be so easy for the subject as to seem almost insulting should be followed immediately by one that will present a real challenge. The task that is evidently too difficult may be withdrawn lightly before it results in actual failure, with the statement: "Never mind about that; here is something else I want you to try." The second task must be easy enough to insure success, with the materials for the first task presented being put quickly out of sight. It may require several trials to find something difficult enough to command the subject's interest and respect yet easy enough for successful achievement. Herein lies the importance of having a varied supply of performance materials, sufficient for coverage of all mental levels.

A carefully prepared case history, containing full information concerning the subject's school record and play interests, is very helpful to the examiner in finding a task of the right degree of difficulty for opening the examination. It happens with inconvenient frequency, however, that a child is presented for examination with no history at all. Perhaps the examiner's only way of preventing some tactless person from stating the essential facts in the child's presence is to close the door and assume full responsibility for finding out by trial and error what tests can be used most appropriately. Even the identity of the subject may be unknown, but the examiner should not begin by inquiring his name and age. This is the time when it is urgent to make the child feel at ease, not the time to remind him that he is with a stranger. The facts may be ascertained casually, later in the examination.

The final task is as important as the opening task. The examination should always be brought to a close with something that will leave the child with a sense of achievement, not a sense of frustration. It is a pleasant ending when the subject completes in triumph the final and most difficult task of a series. In case he reaches his limit in the course of the series, however, the examiner must be prepared to smooth over the failure. Other tasks must be presented, as many as necessary, until something is found which makes a comfortable ending for the examination.

The preliminary task of an examination has been described by some as the "shock absorber" and by others as the "ice breaker." The closing test may be referred to as the "buffer task." Specific suggestions for them will be offered in the chapter on unstandardized tests.

The most important thing for the examiner to remember is that the subject's point of view must be kept in mind at all times. Children differ very widely in their aptitudes and tastes, and they should not be expected to react alike. It is the examiner's task to understand the particular child being interviewed and to adapt the examination to the needs of this individual. If there is anything that the children referred for clinical study share in common, it is that their self-respect has suffered a shock both because of the reasons for referral and also because of the referral itself. Anything in the examination that shakes still further a child's confidence in himself and his own powers may do him more damage than can be counteracted by the course of treatment recommended as a result of the study. On the other hand, an examination which stimulates his sense of achievement may have a constructive value for him quite apart from subsequent treatment.

2. Distribution of Unpleasant Tasks

It is neither possible nor desirable, except for a very young child, for the examination to be treated as a game. The child of school age should take it as seriously as a school task, and should find his satisfaction in intellectual effort and achievement rather than in the personal attention he is receiving or the entertainment afforded by some of the tests.

The examination as a whole should be sufficiently attractive to command the subject's spontaneous cooperation, but this does not mean that all unpleasant tasks are to be omitted. Certain tasks that are essential to the examination may happen to be distasteful to the subject, and almost every test may include items not to his liking. Nothing must be permitted to interfere with the adequacy of the examination; but it is still possible, by careful distribution of unpleasant tasks, to show some respect for the subject's likes and dislikes

and to keep him in a cooperative frame of mind. The arithmetical test, which most children dislike, may be presented sympathetically: "Do you like arithmetic? No? Well, I'm sorry, but please do the best you can with this. I'll have some more puzzles for you later." In the orally-presented test it is important to avoid a series of consecutive failures. Rigid adherence to a given order of presentation is usually unnecessary, but any deviation from standardized procedure requires care on the examiner's part to avoid giving the subject a clue to the answer of some question which is to come later. When the sequence of the questions is unimportant, it is possible to take certain liberties with the arrangement while preserving unaltered the essential difficulty of each question.

3. The Place of Praise in the Examination

It has long been recognized that praise is an important factor in holding a child's attention, and there is perhaps a tendency to use it a little too freely.

Some children are so diffident that they have to be coaxed along all through the examination, requiring considerable praise to keep them responding at all. Sometimes, also, it is the over-confident child who demands praise for every response, merely because he is fond of receiving attention. In general, it is better to reserve the praise for the response or the performance that is really worthy of special notice. The child should not be encouraged to expect praise for every correct response, nor to assume that a response received in silence is necessarily incorrect.

In the orally-presented test it is never permissible to praise an incorrect response.[1] Undeserved praise will either make the subject feel foolish or else have the effect of giving him false information. A better way to restore his confidence after a failure is to pass on quickly to a question that he can answer to his own satisfaction or a task that he can perform successfully. Young children frequently

[1] Many years ago, when examining an adolescent boy who needed all the encouragement that could be given, the writer told him, "That's right." The boy replied under his breath, "No, it isn't." His comment was not intended for the examiner's ears, but it made a significant contribution to her education.

respond well to the tone in which a question is asked and occasionally to a playful change of tone.

Praise should be spontaneous and sincere, not automatic. There are many ways of minimizing a subject's failures and helping him to forget them. It is important to ease the situation for the subject who recognizes his limitations, but this can be done without falsification and without making praise so routine as to destroy its significance.

4. Pencil-and-Paper Tests

There is an important place in the clinic for the verbal-performance test—the task which is essentially dependent upon the child's understanding of verbal instructions but which does not call for an oral response from him. Defective speech is a common reason for a child's referral to the clinic. In some cases the child has a language of his own, understood by those in daily contact with him but almost unintelligible to a stranger. His ability to follow specific directions offers a good index to his understanding of language and his general comprehension. Written tests are useful for the subject who can read, whereas picture-marking tests may be used for illiterate subjects. Tests that have been standardized without time limit are especially at a premium. The test that is complete on one page possesses a great advantage over those in booklet form. Tests having a rather narrow range of discriminative capacity are better suited to individual study than are wide-range tests intended for coverage of a heterogeneous group.

When the examiner is in doubt concerning the subject's ability to read, it is well to begin by requesting him to write his name. While looking carefully in another direction, the examiner may note whether the child writes automatically or laboriously. A short written test may be presented, with the request that the subject read aloud the first line. Further instructions for performing the task may be given the child who thus proves his ability to read; but the one who cannot read should be given a picture test, so that he may still have a use for the pencil. It is important that the picture page be close at hand. The page that the child failed to read may be put out of sight and forgotten.

Children who read with fair facility are likely to prefer the written test to the orally presented one. It is less personal, and therefore less humiliating, to guess at a response or even to omit a line than to face the examiner and acknowledge openly, "I don't know." Of course the written test should not be examined in the presence of the subject, except when it can be done without his knowledge.

5. The Preschool Child

Examination of preschool children is a problem in itself, requiring special materials and special aptitudes on the examiner's part. The adaptation of the examination to the individual child becomes increasingly urgent as we go down in the age scale. Our tools for examination, imperfect for subjects of any age, are notably weak for young children who have never been subjected to school discipline.

Tests standardized in nursery schools have only limited applicability to the child who has been cared for almost exclusively by his mother. A test scored by speed of performance is of little value in the examination of a three-year-old who has his own ideas of what he wishes to do with a given toy. What the stop watch measures is not necessarily the time actually needed by the child for performing the task, but may be rather the time required by the examiner for persuading him to perform it at all.

For a young child, or an older one of low mental level, it may be quite impossible to conduct an examination in strict accordance with rules. It is frequently necessary in practice, however undesirable in principle, to permit the mother to be present. In such a case the mother is requested to hold the child on her lap, with his back toward her, and on no account to speak to him unless her assistance is solicited. This means that there are two persons whose cooperation must be obtained, the mother being perhaps less cooperative than the child. It cannot be expected that an examination thus conducted will give trustworthy results, but the findings may be of some value. There would be no examination at all if the child were screaming for his mother and refusing to take any notice of the test materials.

6. What to Observe in the Subject's Performance

At no time during the examination should the subject be observed so closely as to make him feel unduly self-conscious, and yet the examiner should at all times be keenly alert to everything that the subject does. It is an important part of the examiner's task to see more than he seems to see. While the subject is working on a written test or a performance task, the examiner should appear to be busy with some desk work. When the subject looks up to see how closely he is being watched, he should find the examiner rearranging the contents of a drawer or looking up something in a card index. Reading matter should be available for intervals when there is nothing more active to be done. However, something requiring use of the hands is always to be preferred to mere eye work. The subject feels more relaxed when the examiner is obviously doing something not related to the examination.

While the examiner is pretending to be busy, very little should actually escape his notice. By keeping his eyes open to the little things that make this subject's performance distinctive, he can learn much that is not included in the score.

Anything atypical about the child's physical development should be noted, especially any physical defect that might place him at a disadvantage. A speech defect or impediment is of great significance, because it might invalidate the child's score in an orally-presented test. Defective coordination might have serious effects in performance tests scored by speed. Are the child's movements jerky, or do his hands tremble?

What is the child's attitude toward the examiner? Is he friendly and responsive, shy and withdrawing, or definitely hostile? Is he curious and inquisitive? Does he snatch at the stop watch and meddle with everything within reach? Is his behavior that of a well-trained child, or is he so uninhibited as to be seriously troublesome? Does he try to be helpful in putting materials back into their boxes? Is he alert and interested, or does he behave as if bored?

Does the subject use his two hands coordinately, or is there some interference between them? Does he work lazily with one

hand, while using the other hand as a support for his head? Does he let the paper slide around on the table when he touches it with the pencil, and does he straighten it with his elbow? Does he push the formboard out of place with each placement of a block? Does he respond to the increasing difficulty of a formboard series and show greater interest as the tasks make heavier demands upon his attention? If so, at what stage in the series does he hold up his head and work with both hands? Does he become easily discouraged by the difficulty of a task and say that he can't do it or that it can't be done? Is he irritated by the difficulty, and does he show it by becoming rough with the materials? When inserting a block which fits too tightly to slide into its place on a slant, does he try to force it by pounding with his fist, or does he take care to hold the block parallel with the board? How well does he follow specific instructions in a series which is preceded by a demonstration task? In the color cube series, does he frequently remove a block that is correctly placed? Does his comprehension of the essential principle underlying the task show improvement with practice? Does he learn to rotate the red-white or blue-yellow block consistently until he finds its correct position, or does he try two or three incorrect placements and then repeat these errors by reversing the direction of the rotation?

Manipulative tests afford better opportunity for observation than do orally presented language tests. The performance of a difficult task allows time for the examiner to record what he observes. The subject, having his eyes and attention on a table task, is not so keenly aware of the examiner's note-taking as when the two are facing each other squarely. What the subject does at any time, however, should be noticed and remembered, even if it has to be recorded after the interview. Almost nothing that the subject does is so trivial as to be unworthy of notice.

7. Physical Equipment of the Examining Office

The furniture of the examining room need not be costly, but it should be suited to the purpose even when this necessitates having it made to order.

One important piece not easily found on the market is a cabinet, easily movable but usually standing at the examiner's left. Some of the shelves may be spaced about one inch apart, for holding such materials as small formboards and folders containing test forms. It can be large enough to contain everything needed for a routine examination, without being too heavy to be pushed around freely with one hand. This cabinet should be closed on three sides. During an examination it should stand at such an angle as to keep its open front entirely beyond the subject's range of vision. The child must not be permitted to know the number of tasks in a series, because he may reach his limit long before the series is completed. In order that the last task presented may be for him the end of the series, all additional materials must be safely concealed from him. A cabinet that makes this possible can be built at moderate cost.

The examining table should also be built to order, with a draw-shelf for the examiner's use while taking a record. The subject should not be permitted to see his record, except for some special reason. Occasionally when a three-trial practice series results in a perfect learning curve, it is encouraging and stimulating to the subject to be shown the record. More often it is important to keep the record out of sight, and it is equally important to avoid letting the subject see that it is being hidden from him. A sliding shelf in the center of examiner's side of the table, about three inches below the top, makes it possible to conceal the record so easily and naturally that it is almost as if no record were being taken. A record written in full view of the subject may excite his curiosity, whereas he is hardly aware of a record taken under cover.

An additional advantage of the shelf is that it makes the table adjustable in width. When the examiner is handing out materials, it is convenient to be able to reach across to the subject's side of the table. But while the subject is working with the materials, the examiner should be able to move back far enough to leave plenty of foot space for two pairs of feet. By pushing the shelf in when presenting materials and drawing it out for recording, the examiner can have the table of the correct width at any given moment.

If the table can be built as a desk, with drawers on each side of the knee hole and a shallow drawer above the shelf, so much the

better. Even without drawers it should be preferred to a desk of conventional design. The subject cannot work comfortably at a desk having a closed back and allowing no space for his knees.

Fiber board is recommended for the top of the table or desk, because it has a good surface both for writing and for handling blocks. It is smoother than a wooden top, harder than linoleum, and less noisy than plate glass.

It is worth while, if possible, to have a "junior" dining chair for the use of small children, including a foot rest.

For a clinic set up in temporary quarters it may be impossible to design special furniture, but there are many ways of improvising crude facilities that will serve some part of the purpose. The recording may be done in the examiner's lap, on a small writing board with a clip at the top. This is an uncomfortable and too-obvious way of keeping the record out of sight, but it is better to take the record in this way than to take it in full view of the subject. Also, with the aid of gummed paper, one can build a receptacle for performance materials out of cardboard salvaged from discarded cartons. This box can stand in a chair at examiner's left, serving passably to keep the test apparatus from being seen prematurely by the subject.

The effort required for making and using these things has a value of its own, in reminding the examiner that no detail of technique is too trifling to receive careful attention. When the validity of an examination is artificially damaged, it should be by reason of some factor beyond the examiner's control. This will happen often enough at best.

CHAPTER III

SOURCES OF ERROR IN USE OF TESTS

1. Statistical Fallacy

There is a fundamental fallacy in the use of statistical criteria for evaluation of individual achievement.

The actuary can determine by statistical methods what is a safe gamble for life expectancy at the age of forty, but he does not presume to predict the exact year in which any particular man of forty will die. Nor has he any interest in the application of his statistical tables to the life expectancy of any individual.

Norms for mental tests are established by the same methods as actuarial tables, based upon a larger number of items and a smaller number of cases. It is safe to assume that the norms hold for the majority of the subjects to whose achievement they are applied, but they do not necessarily hold for any particular individual. There are known to be exceptions among normal subjects. In any group of children examined for the purpose of standardizing a test, even when they are selected as being well-matched for school achievement, it is expected that the scores at the ends of the range will deviate markedly from the average score. The group includes a few exceptional children whose capacity as indicated by school achievement is not fairly measured by the test used in the examination. It naturally follows that the particular child studied in the clinic by means of this test may happen to be one of the exceptions.

The validity of the statistical method for group study varies directly with the number included in the group, but it is not primarily for group study that the mental test is standardized and used. Its practical significance for individual study in the clinic begins at the point where its statistical validity breaks down.

2. Constant Error in Clinical Use of Norms

Most tests are standardized by being presented to school children during school hours. Anyone who has standardized a group test in the elementary grades must have observed the immediate responsiveness of the children when they are instructed to put away their books and give their attention to a visitor who has just entered the room with a pile of papers. Anything unusual introduced into the day's work is regarded as a special treat. The children are alert, eager with curiosity, so impatient to read the test forms that it is difficult to induce them to wait for the signal to start. The test presented individually elicits a similar response. It is a special privilege to be sent to the teachers' room for an interview with a visitor, and each child enters the room with an air of expectancy. Whether for a group or an individual examination, the half-hour release from routine class work is at a premium. It is easy to obtain excellent cooperation. What is measured by the test in the collection of data for the establishment of norms is something very close to maximal achievement.

The typical clinical interview presents a contrast that is almost painful. In most cases children are referred to the clinic because of some deficiency either in school work or in conduct. The child of school age usually feels it very keenly that he should be singled out for special examination. Consequently, he enters the room with resentment and is in no mood to do the best of which he is capable. To be sure, the tactful examiner can usually overcome his hostility and arouse some interest on his part; but it is exceptional when he brings to the examination the alert spontaneity of the normal child examined in school.

The clinical examination cannot fairly be assumed to measure what is measured for normal children except when the conditions of the clinic are as favorable to maximal achievement as were the conditions under which the tests were standardized. In the clinic, any disturbing factor has a tendency to lower the subject's score. The sudden noise of a passing truck may drown out the barely-audible response of a timid child, and he may lack the courage to repeat it after the truck has passed. The examiner, not knowing what the response was, has to put it on record as failure.

Occasionally the scale or battery may be weighted in favor of some child whose experience and training have been exceptional. However, inasmuch as children studied in the clinic frequently come from under-privileged homes, the child's opportunities for the experience measured by the test are more likely to have been sub-average. This is not a constant error in the sense of being universally applicable, but it applies to a large proportion of cases. At best there is a tendency for the clinical subject's score to fall short of his potential achievement.

3. Interference of Emotional Factors

Apart from the emotional response which the examination may arouse, there is frequently a basically unhealthy emotional state associated with the reason for referral. A child who is doing well in school and who is happy in his home life is not often sent to the clinic. It is usually the maladjusted child who needs what the clinic can offer. Whatever the type of maladjustment or the immediate reason for referral, the subject may feel the situation more keenly than he cares to show. When a child goes through the examination with perfunctory obedience and with no trace of a responsive smile, it is well to consider the possibility that his attention may be centered upon something painful. On no account should there be any probing at this time, if at all. The child's reticence should be respected, but his lack of responsivenessness is something for the examiner to keep in mind when reporting the case. The child's attitude may be wholly unexplained, but it should at least be recognized as having possible significance.

The subject's emotional reaction to the examination itself is frequently all too apparent, especially in response to those questions or tasks that are too easy or too difficult to be appropriate to his mental level.

It is not possible to know and not permissible to guess to what extent a child's achievement may be affected by his emotional reactions; but it is safe to assume that the effect, if any, is to lower his score. Among the various factors tending to produce a constant error, this is perhaps the most difficult to evaluate.

4. Erroneous Estimate of the Subject's Cooperation

A child may seem to be doing his best when actually his attention is on something far removed from the test situation. Also, a child may actually be doing the best he can while seeming to be inattentive or careless. Startling errors in the examiner's estimate of the subject's cooperation have been brought to light, when someone challenged the results and requested a re-examination. Even the most careful observer may occasionally be deceived by appearances.

The danger of misinterpretation was forced upon the writer's notice in 1928, in the examination of a young woman who had recently been admitted to a state hospital.

The patient responded very hesitatingly to some of the Stanford-Binet questions, frequently moving her lips without producing a sound. There was no open resistance to the examination, nor anything to suggest negativism. Her manner was friendly at all times, and she seemed to be making a desperate effort to comply with every request. She smiled responsively when praised for a correct answer.

When a written test was presented her behavior was similar to that observed in a retarded child who will not acknowledge that he cannot read. She studied the page long and carefully, as if trying to read, but could not be induced to pronounce a word or to use the pencil. She did not refuse, but continued looking at the printed page until it was finally taken from her. It was painful to witness her apparent effort to cooperate in the examination.

The examiner reported a rating of about eight years, adding that the patient's cooperation was satisfactory; but the physician challenged the findings, and persuaded both patient and examiner to try again. At the second examination the patient earned a twelve-year rating by a written battery. Her explanation of her behavior at the earlier interview was that she "didn't feel like trying." Actually she had been trying not to try. The effort was evident, but what it represented was not even suspected by an experienced examiner.

When it is possible to have a child brought to the clinic for another interview, it is desirable to have the second examination conducted by a different person and with a different battery of tests. The second examiner should be informed what tests were used and what others are recommended for the second trial, but should not

see the record of the first examination. Beyond knowing that the child is one calling for unusual delicacy of approach, the second examiner should take up the case with the minimum of bias.

5. Fatigue of the Examiner

It is only by putting all one's personal presence into the presentation of each item that one can obtain good cooperation from the child in the clinic. It is, at best, very exhausting work for the examiner.

Standardization of a new test is far more stimulating, both to the author and to the student-assistant, than is the routine use of a test after it has been standardized. The examiner in a standardization project can maintain a high level of interest during the entire working day. He is carried along by the alertness of the children, by his enjoyment of each response that is new or distinctive, and by the satisfaction of seeing the records accumulate. Fatigue may be felt at the end of the day, but it is scarcely noticed while the work is in progress. The examiner's interest in his task is spontaneous, not forced. He watches the clock to see if there is going to be time to take another record, not to see if it is nearly time to go home.

But the clinical use of the same tests day after day becomes deadly monotonous. After the technique of presenting a test has been mastered, continued experience does not make the work easier. No short cuts are permissible in the use of a test evaluated by norms. The examiner may feel sure that a given question will, or will not, be answered correctly, but no feeling of certainty on his part justifies him in recording a response not actually received. Habitual use of any test brings its defects more and more into focus, and the examiner finds it increasingly irritating to abide by the rules for presentation—rules which he could greatly improve in the light of his own experience.

Many other unavoidable factors contribute to the strain under which the examiner works. One child may stubbornly refuse to respond, resisting all the blandishments of the diplomatic examiner; another may treat the situation as a game planned for his entertainment, occasionally giving facetious responses; and a third child may

take the examination so seriously that it is distressing to offer him any task which may result in failure. At times it taxes the examiner's resources heavily to keep the stop watch beyond the reach of a child's meddlesome fingers.

Because of the various factors that make the examination difficult, there is a limit to the number of children who can be properly and adequately examined by one person in one day. In many clinics the demands of a day's work force the examiner, more or less habitually, to continue taking additional cases after he is no longer fit for work that is so exacting. A spiritless examiner cannot inspire a subject to put forth maximal effort. The examiner's fatigue should be counted as one of the most common sources of error in the findings. No one knows how many times it has served to lower the level of the subject's achievement.

6. The Subjective Factor in Test Presentation

Too much has been claimed for the objectivity of the psychological examination. The conditions under which a test is presented can be controlled only to a limited degree, and the conditions under which it is received are still less subject to control. While we are struggling to make the presentation as objective as possible, the subjective factor creeps in by the back door.

Furthermore, it is only within limits that objectivity as such has value. A group examination is more nearly objective than an individual one, but we do not for that reason consider the group examination the more trustworthy. The Binet Scale could be presented more uniformly and more objectively by a phonograph record than by a living person, but no one would expect the subject to react normally to so unnatural a situation. It is recognized that strict control of conditions could be carried so far as to introduce more errors than it would eliminate. Consequently, we do not go quite to the limit of attainable objectivity, but we do go far enough in that direction to sacrifice something of value. While our attention is focused upon such exact determinations as can be wrung from the examination, we neglect to take full advantage of the opportunity it offers for studying the reactions of the subject for what we can learn from them.

It is important to keep clearly in mind the distinction between the objective and subjective factors. Otherwise we may deceive ourselves concerning the measure of objectivity that is attained and the kind of objectivity that is worthy of attainment.

7. Estimate of "Intelligence Quotient"

It is generally recognized that mental development is more rapid in early childhood than during adolescence. On the other hand, the current method of converting the "mental-age" into the "intelligence quotient" is based upon the assumption that mental development proceeds at a constant rate up to the age of fourteen, if not fifteen or sixteen. Dividing the mental age by the life age to obtain what is called the intelligence quotient does, in effect, hold the child of fourteen responsible for developing as rapidly as the child of five years. In cases of retardation, it discriminates against the older child as compared with the younger one.[1]

When a child's mental-age happens to agree with his life age, it makes no difference how the findings are expressed, inasmuch as the intelligence quotient—by whatever method computed—will be very close to 100. But the wider the discrepancy between the life age and the mental-age, the more inaccurate is the intelligence quotient as usually computed. This is true for a deviation in either direction. It explains why a very high intelligence quotient is so frequently found among preschool children, and it accounts also in many instances for the unexpectedly low intelligence quotient found at the adolescent level. The figure that is commonly believed to be a percentage of normal mental capacity has different significance at different ages. This is illustrated in exaggerated form by the case of a mentally retarded boy who grew up in a state training school. He entered the school in his seventh year and was examined

[1] See Wechsler (155, chapter 3). The essential fallacy involved in the current method of deriving the intelligence quotient from the mental-age, as introduced by Terman in 1916 (143), has been well covered by Wechsler with reference to mental impairment in later life. It is significant here in that this fallacy applies also to the developmental period. As Wechsler has shown, the course of change is not linear. There is no short-cut path to a rating which indicates a true percentage of average normal achievement as compared with persons of the subject's own age. The only way to obtain a rating which really means what the percentage form implies is to standardize the test age-for-age, as Wechsler has done.

six times during a ten-year period. The series of intelligence quotient estimates was as follows: 66, 62, 52, 50, 49, 44. During the same period his mental-age showed a gain of thirty months, which is better than would be expected if the progressive loss in intelligence quotient were to be explained as due to deterioration. For the clinical subject whose intelligence quotient is put on permanent record as a result of a single examination, too much depends upon the age at which he happens to be referred to the clinic.

It is undeniably convenient, in the study of children in the developmental period, to have the results expressed in the form of a percentage; but there is at best quite enough possibility of error in reporting individual achievement in terms of any statistical data, without introducing an additional error. Until tests can be standardized on such a scale as to have independent norms for each age covered by the test, so that a child's score may be compared with the achievement of other children of his own age, there is no justification for forcing his rating into the form of a percentage. The mental-age rating may not be correct in a given case, but it can be offered honestly as meaning what it is understood to mean. It is none too trustworthy, but at least it does not involve any actual distortion. The physician and probation officer should learn to think in terms of the mental-age, as was customary before the intelligence quotient came into general use.

8. Other Possible Errors

There is almost no limit to the variety of technical errors which can occur. No one is quite infallible, even in so comparatively simple a matter as counting the points of a score. Unfortunately, there are many possible errors not so easily checked.

It seems almost unbelievable that a psychologist's failure to inquire how long a child had been wearing glasses could be responsible for the recommendation that he be placed in the state school for mentally deficient children, but this is one of the things that can happen. The child was a boy of thirteen who had just completed third grade in school. The case history was unusually full and adequate. It was known that the boy had been motherless from

infancy, that his father was a heavy drinker, that he had been passed around among relatives who did not want him, and that he had never remained in any given school for more than one year. The child's history was a story of flagrant neglect, but it was not known—because it did not occur to psychologist or physician to inquire—that his first four school years had been almost wasted by reason of his defective vision. Nor was it recognized that he had made normal progress in school after his vision had been corrected. His education had not actually started until he reached the age of ten, and he had covered more ground in three years than would be possible for a mentally deficient child. And yet, all that saved him from being committed to the training school was that there was no vacancy. It was necessary to make a temporary disposition of his case, by reason of which he was accidentally brought to the writer's notice.

9. Practical Effects of the Various Inaccuracies

It is because the psychometric test is so widely accepted as an "instrument of precision" that these possible errors become a source of real danger. The mental test holds almost unlimited possibilities for usefulness in the understanding of a child, if used intelligently; but let any test or system of tests be widely recognized as the "accepted standard," and it will be found to contain vast potentialities for mischief and injustice.

In some states it is required by law that a child's intelligence quotient be the determining factor with reference to his commitment to a state school or his admission to the ungraded class in the public schools. Such laws were presumably adopted for the protection of children, on the assumption that these measures would prevent a near-normal child from being institutionalized. The laws could serve that purpose, however, only if the test findings were as infallible as the state legislators believed them to be. So long as it is possible for a child's rating by the test to be far below his actual capacity—whether because of poor cooperation, physical defect, inadequate educational opportunities, or for whatever reason—the psychological examination does not afford him proper protection.

No criterion based upon statistical findings can safely be accepted at face value for a purpose so important as the institutionalization of a child. Least of all should the intelligence quotient, which is the most untrustworthy figure included in the record, be thus accepted.

It is only by many years' experience in the personal use of tests that one learns to spot the possibilities of error, and yet the findings are being used for disposition of cases by persons who are wholly uninformed concerning possible inaccuracies. In some social agencies it is considered desirable to present in the annual report a table in which all children served by the agency are classified according to intelligence quotient. In a home for young unmarried mothers a girl was examined at her baby's feeding time, and the room used for the examination communicated with the nursery where the child was screaming. The mother's inability to give attention to the test was a better index to her mental capacity than was the rating she achieved, but the superintendent of the Home was indignant when a report on the girl's intelligence quotient was refused. In some state institutions the demand for a numerical rating is so absolute that the psychologist can refuse it only at the risk of his position.

In the pathological laboratory the "doubtful reaction" has a recognized place, and a new specimen is obtained for further examination. In the psychological clinic, however, it is not always possible to bring the child again for re-examination; nor is there any guarantee that he will give better cooperation in the second interview than in the first. In an examination which involves the personal equation of the subject, it may not be possible to obtain results which can safely be used for disposition of the case.

The uncritical acceptance of test findings on the part of the general public dates from World War I, when psychological service enjoyed a very unhealthy boom. The Stanford-Binet scale (143) rode into public favor as the accepted standard, on a tidal wave which has not fully receded after a generation of continued development of tests. In many state institutions the use of this scale, in the original or revised form, is required by departmental regulations which are as binding as state laws.

The responsibility for this rests with psychologists. We have come forward with an offer to furnish intelligence ratings at wholesale rates. We have led the public to overestimate the degree of accuracy with which native mental capacity can be determined. We have claimed too much for our tests and have been taken at our word. The situation is essentially one of our own making.

CHAPTER IV

THE COMPOSITE SCALE VERSUS THE BATTERY

1. Explanation of Terms

The term "composite" is used here as referring to a scale that is scored as a whole, resulting in a single score or rating; and the term "battery" is reserved for the series of tests or sub-tests, each of which is so standardized as to yield an independent rating. It is essential that the sub-tests of a battery have a common denominator, usually the mental-age rating; and the end-result of the battery is usually expressed as a median rating.

The best known and most widely used composite scale is of course the **Stanford-Binet Scale** (143). In the same class, although not so generally recognized, is the scale known as **Kuhlmann Tests of Mental Development** (97).

Having already commented at some length upon these scales (79, 22), the writer has nothing to add at this time.

The most complete scale known to the writer which is standardized in the form of a battery is the **Detroit Tests of Learning Aptitude** (10). This scale will receive further notice in a later chapter. The **Kuhlmann-Anderson Intelligence Tests** (97) are also in the form of a battery.

A test or scale may be standardized both as a composite and also as a battery of independent units. For example, the **Detroit Primary Intelligence Test** (9) carries three sets of norms: one set for the entire booklet of seven one-page tests; another set for each of the seven tests treated as a unit; and still another set for certain combinations of two sub-tests. One who uses the scale in an individual study may use it in whole or in part, according to the method best suited to the subject.

In a sense, any test consisting of more than one item is of necessity a composite scale. It is difficult to devise two items which are

similar in kind, or to insure equal score value for them except by artificial weighting. Even in a formboard, each figure has its own distinctive form and presents its own distinctive difficulties. It is not easy, if it be possible, to find a task that can be graded in difficulty quantitatively without being affected qualitatively. The maze, for example, permits gradation in size without obvious deviation from a given pattern; but two mazes of the same size and pattern may be found empirically to differ appreciably in difficulty. The writer has been trying for thirty years to devise gradable tasks in both language and performance tests but has reached the conclusion that the graded series is something to be only approached, not attained.

2. Weakness of the Composite System, in the Clinic

It is the composite examination, rather than the composite scale, that is open to criticism as a means of measuring individual achievement.

The inherent weakness of the composite examination rests upon the fact that a child's cooperation is selective. Personal tastes differ, and no particular test item has universal appeal. In the writer's experience, the formboard task (of suitable difficulty for the subject) is the test that arouses the keenest interest and is on the whole the most valuable means of diverting a subject's attention from the unpleasant incident that caused him to be referred to the clinic; but the formboard task is also the test which elicits the most violent resentment on the part of a subject who has a special antipathy for puzzles. So also the repetition of digits or sentences is a rather attractive task for the subject who has a good rote memory and likes to show it off, whereas it is highly offensive to the subject whose memory does not function in that way. The clinical subject cannot be counted upon to do his best work in a task that he dislikes; and it is unreasonable to expect a scale containing sufficient material for an examination to be acceptable in its entirety to all the subjects for whom it is to be used. Almost of necessity, some items will be so offensive to some subjects as to fall short of commanding their full cooperation.

The composite examination allows no room for recognition of individual differences or individual tastes; and the scale which is intended to serve as the whole examination is usually too time-consuming to leave any time for supplementary tests. When the examination consists of a single scale, there is no check on the results.

It was in 1915, when the Goddard-Binet scale (47) was the only standardized test in common use, that the writer first felt the need of a flexible system as opposed to the rigid scale which yields a score only when used as a whole. This need was not met even provisionally for more than ten years. Various group-standardized school tests were tried out for use in clinical studies, and the Terman Group Test (145) was selected as being well suited to individual presentation. Because of its division into ten sub-tests, having time limits of not over four minutes, it permits a moment of relaxation between tasks; and it is therefore a comfortable form of examination, as compared with the single-task written test which keeps the subject working at a breathless rate for upwards of twenty minutes. The Terman Group Test, in the 1920 edition, is still used by the writer in examination of a high school student who has reasons of his own for wishing to earn a high score; but it is a weak test for an unwilling subject, because of the fact that it can be scored only as a ten-test composite. Its limitations came to light in 1922, when it was used for a girl who had refused to answer the Binet questions but consented to try some written work. She gave acceptable cooperation on the first four pages and on some additional sub-tests, but refused to have anything to do with the two arithmetical tests. It was learned subsequently, by indirect means, that she possessed more ability in arithmetic than one would infer from her reaction to these tests. Her unwillingness may have been due to lack of confidence, but not to total incapacity. Her scores in these two tests might have been lower than those she achieved in the other sub-tests, but they would have been better than zero. The examination was invalidated, and the materials at hand did not include any test to which she could be prevailed upon to submit in full. It was this incident, more than any other clearly-remembered examination of that period, which led the writer to spend twenty-five years developing short, independent units.

Furthermore, the unsuitability of some items for some subjects is not merely a matter of taste. One of the problems most frequently referred to the clinic is that of the child who reaches reading age without having learned to read. His achievement in any language

test will be somewhat lowered by his inability to read, because the norms are based on the achievement of children who can read (p. 4). If his score be lowered still further by the use of a scale containing items that pre-suppose reading knowledge and will necessarily be scored as failure, there is danger that he may be diagnosed as a case of mental retardation without having had a fair examination. Someone may assume that he is incapable of learning to read, when in reality he is a promising subject for special effort in remedial reading.

Also, the child having a speech defect is severely handicapped in a scale that makes heavy demands upon speech. No questions calling for oral response are suitable for him, unless they admit of being answered in a few words that can be easily understood.

There is a place for written work in the examination of the subject who can read, and for the involved question in the examination of the subject who talks freely and easily. Test material of various types may be very useful for some subjects while being wholly unsuitable for others. It is the inflexibility of the composite examination that impairs its usefulness in the clinic rather than the weakness of any particular item.

3. Strength and Weakness of the Battery

The essential feature of the battery is its flexibility. The examination can be adapted to the individual subject. Each unit may be selected as being reasonably close to the mental level indicated by the preliminary "shock-absorber" tasks. Guided by the subject's reaction to each task presented, the examiner can avoid using any test that is grossly inappropriate, and can keep at a minimum the nondiscriminative matter which is wasteful of time and needlessly annoying to the subject. Any test in which his cooperation is unsatisfactory may be thrown out, and the rating may be based upon those tests which are most representative of what he can do rather than what he elects to do.

Another advantage of the battery over the composite examination is that any unit which becomes obsolete may be discarded, without invalidating the norms of the whole system.

The test that is worthy of a permanent place in the progressive clinic is the rare exception. Language tests are likely to lose something of their significance with the passage of time. Unfortunately for the long-range usefulness of a language test, the item related to current events is especially significant at the right time, but practically useless when out of date.[1]

The nonverbal test does not become outdated in quite the same way as a language test, but it loses something of its value when someone offers a better or cheaper or more convenient test that will serve the same general purpose.

The battery adopted by any particular examiner should be in a state of flux, undergoing change and improvement from year to year. Any new test that becomes available may be introduced into the system, enlarging the collection of tests from which the examiner may draw the material for examination of a given subject, or superseding some test that has not fulfilled its promise of serviceability. The system admits of almost unlimited expansion. It is conducive to progress in the development of test units, because it encourages small contributions from students who have not the resources for undertaking a heavy standardization project. It makes possible the development of an instrument which is the product of many persons of diverse training and habits of thought, as opposed to the product of a small group of students dominated by a single mind. The comparative ease with which a battery can be modified—first with reference to the examination of the individual subject but more importantly with reference to the equipment of the individual examiner—may have far-reaching consequences for the clinical examination of the future.

The most serious weakness of the battery recognized by the writer is the shortage of test material standardized in the form of

[1] The "stamp question" in Kent Emergency Test (76) serves to illustrate the unexpected way in which an item may become obsolete. "Whose picture is on a 2-cent stamp?" was a good question in 1925, when the test was formulated; but after letter postage had been raised to three cents, the question became entirely unfair. After receiving numerous complaints, for a period of several years and from all over the country, the writer revised the test by omitting the questions that for various reasons had been found troublesome. The revision made the test too short, but improved its quality. It was only by accident that material was available for standardizing the revised scale; and without this material it would have been necessary to discard the whole test.

independent units. It is not easy to devise questions or items of any kind that are fundamentally sound and serviceable. A great many of the best items yet devised are already used in elaborate composite scales, in which they can form no part of a flexible system. There are hundreds of written tests prepared for classification of school children, but it is not easy to find among them any which meet all the requirements of a clinical examination. Most of them are published in booklet form and standardized as a whole, usually requiring more time than can be allowed for a single unit of a battery. The clinical examiner who prefers a flexible examination to a fixed scale is hard pressed for units short enough to be used in a series and yet long enough to yield an acceptable rating.

Another acknowledged weakness of the battery is that the various units do not have equal value. They may vary widely in length, complexity, range of coverage, and adequacy of standardization, to say nothing of the inherent differences which make one item intrinsically more significant than another. When making up a series of mental-age ratings, it does not seem fair to let a task occupying only a few seconds have the same weight in determining the median as the task requiring several minutes.

The most fundamental differences in value are of course not peculiar to the battery. For the composite scale, as for the battery, two items that carry the same score value may differ greatly in real significance. They can be made statistically equal by artificial weighting, but this does not affect their true value as to what they reveal of a child's capacity. Nor does the profile chart, showing the comparative ratings by different tasks, throw any light on the relative importance of the tasks. The correlation of a given task with a larger scale offers an index to its significance in the large, but does not indicate its value for any given individual. The composite scale has an advantage over the battery in respect to uniformity of standardization, but not with reference to differences in the inherent significance of the various items.

4. Structure of Performance Scales

We are indebted to Pintner and Paterson (121) (who quite possibly may have builded better than they knew) for the first battery

of independently standardized units. This method was given us in 1917; and although the system has not been widely recognized for its distinctive value nor adopted by the majority of test makers, this writer counts it as the most important contribution to the field of mental tests made by anyone since the publication of the original Binet-Simon scale (12).

Accident appears to have played a part in the development of the Pintner-Paterson system. It was certainly a factor in the writer's discovery of its unique value—the accident of being supplied with about half the materials for the series instead of a full set. Although the complete outfit was greatly desired, it was soon found that a shorter series could be made up from the available materials. A little later it was found that any other independently standardized performance test could be given a place in the series.

It is the structure of the scale that deserves so high a place among clinical methods—not any paritcular unit in the series. Some of the sub-tests have fallen into disuse because of inherent weakness. Still others are no longer used in the form originally offered, but have been re-standardized with a new scoring system. The Pintner-Paterson scale was the pioneer effort in the standardization of a test for deaf children. It should not be expected that a system developed more than thirty years ago would still be found serviceable in its original form. The details of the scale, however, are of little importance as compared with the principle of its construction. The feature of permanent value is the form in which the norms were offered, permitting any examiner to select each task on the strength of its suitability for the particular subject being examined. There is no limit to the expansibility of the system. As often as a new performance test becomes available, it may be introduced into the scale to supersede some sub-test that is not giving satisfactory service. The system is at all times in a state of being improved and rejuvenated.

The first sub-test of the Pintner-Paterson scale to be discarded by the writer was the Knox "Cube Test" (118). This task calls for a certain type of visual memory on the part of both subject and examiner, possibly somewhat analogous to the type of auditory memory required for repeat-

ing digits backward. For the examiner, however, there is an important difference between the two tasks: the examiner presenting the digit test need not be able himself to reverse the line of digits, because he can keep his eye on the page while listening to the subject's response; whereas in the Cube test, the examiner is acting as subject while observing the subject's response. The evaluation is subjective, and there is no check on it. Some examiners are able to take a record that satisfies them, as the writer is not. It appears that the ability to follow the pattern of movement with one's eyes is a special aptitude which either subject or examiner may or may not possess.

Of course, the examination would be useless if it measured only those aptitudes in which the subject is strong. It is not unfair to the subject to include a task in which he is notably weak as compared with other persons of his mental level; but it is flagrantly unfair to the subject to have his performance passed upon by an examiner who is unable to "pass" the test from either side of the table. Whatever the achievement of the subject in the Cube test, this examiner cannot qualify for presenting it.

Herein lies the distinctive and fundamental strength of the Pintner-Paterson construction of a scale, as compared with the composite scale: the examiner may omit the task for which he is unable to obtain a satisfactory record, the task for which the apparatus is missing, or the task for which the time record is accidentally lost. He can take the median rating by the tasks that are properly recorded and need not resort to any such mathematical correction as is proposed by Arthur (7) for salvaging a damaged record.

Another feature for which we are indebted to Pintner and Paterson is the completeness with which they published their findings, thereby permitting any student to use the data according to his own best judgment. Four different sets of norms are offered, so that the results of a given examination may be evaluated by the year scale, the mental-age, the point-scale, or the percentile method.[1]

Again, Pintner and Paterson have presented a special table of norms for each variable measured by any test. Seven of the sub-tests yield two records, one for time of performance and one for a count

[1] The writer agrees with these authors in preferring the percentile method as being theoretically the most sound method of evaluation, but has adopted the mental-age on the ground that it is more commonly used. The battery form of a scale can be used only for such sub-tests as have a common denominator, and there are not enough tests having percentile norms to make up a series. Inasmuch as most tests are standardized for mental-age ratings, it has seemed necessary to use this form of standardization.

of "errors." Evaluation of errors is of necessity more subjective than is the time record, and some examiners prefer to base the score upon the more objective findings. This is optional for the examiner. The authors did not attempt to reduce the variables to a single category, but published independent norms for each. The examiner who can establish to his own satisfaction his criteria for counting the errors has the privilege of using both scores, whereas the one who doubts his ability to make an accurate count may disregard the errors and use only the time record. In every possible way, Pintner and Paterson have presented their data so as to permit the utmost freedom in the use of the tests. The material is made accessible to any student who wishes to experiment with it, in form far more convenient than the original records.

Inasmuch as formboards are highly susceptible to the influence of chance, one might reasonably prefer to use them in pairs rather than as independent tasks. The student who wishes to combine the Casuist board and the Two-Figure board, for instance, can find in tabular form all the data needed for experimentation (121, 1921 edition, tables 6 and 8, pages 109 and 112). It requires only a few minutes to express the two series in a single graph and to find that they show a smoother age curve for the two tasks than for either task by itself. It can be seen at a glance that the discriminative range for the pair is from seven to eleven years and that it would be a poor investment of time to use the boards for levels outside this range.

This flexibility contrasts sharply with the rigidity of the composite performance scales standardized by Arthur (7), by Stutsman (140), and by Cornell and Coxe (29). Seven sub-tests from the Pintner-Paterson series have been used in at least one of these scales, and Pintner's "Manikin" is included in all three.

Stutsman's data obtained from preschool children were first published in the form of decile tables (139), from which anyone can derive norms for each separate task. This method of publication has one important advantage over the norms prepared by the author and offered ready-made for the use of others: the student who has graphed the data and smoothed the age-curve for himself knows exactly what has been done by way of artificial alteration, and can

decide for himself which tests can properly be used as independent units.

It was in response to requests that Stutsman later re-worked the data into a composite point scale. The norms were first given out in mimeographed form with the test materials, and were later published in 1931. It is understandable that workers in nursery schools should prefer a single composite score to a loose collection of independent units. They are acquainted with the children in advance of the examination, can have as many interviews as are needed, and can conduct the examinations under standard conditions. In the clinic, however, the examiner has none of these advantages. The full scale is too long and too fatiguing to be presented in a single interview, and there is always a possibility that the appointment for a second visit to the clinic may be broken. The child may not respond well to a strange examiner; consequently the results may be invalid for certain tasks. For clinical use it is advantageous to have separate norms for each sub-test. The writer has never had an opportunity to use the Stutsman scale in full but for many years has used selected tasks, evaluated by norms taken directly from the decile tables.

Neither the Arthur scale nor the Cornell-Coxe scale can be used to advantage in any clinic in which the writer has worked. The Arthur scale is so time-consuming that it leaves no time for language tests.

There is naturally a demand for the composite scale, on the ground that it can be presented more or less mechanically by a psychometrist of comparatively little training. It calls for more judgment on the examiner's part to select sub-tests suitable for a given subject than it requires to run through a rigidly prescribed series of tasks. It must not be forgotten, however, that the results of the mechanically-presented examination are frequently misleading.

The worst feature of the composite scale is that it allows no room for growth. Although at the time of its publication it may be the best available test, there is a possibility that something better may appear within a decade. Owing both to the cost of materials and to the habit-forming propensity of examiners, there is a strong tendency in almost any clinic to continue the use of something

already owned. The apparatus for a performance scale is too costly to be lightly discarded, and the scale may hold its place for many years after other tests have become available. The routine use of the same scale year after year encourages all the workers to pin their faith to statistical concepts.

A more progressive spirit pervades the clinic in which no fixed schedule is habitually followed. Materials for performance tests are acquired piece by piece, and the students may be moved to try their hands at home manufacture. An experimental attitude prevails, and everyone is on the watch for some new test which may contribute to the adequacy of the examination.

The sub-tests of the Pintner-Paterson series are either obsolete or obsolescent, and those that remain in use will probably be superseded within a generation. But the plan of standardization is of such fundamental value that it may be expected to outlive all the composite scales in current use.

CHAPTER V

LANGUAGE AND PERFORMANCE TESTS, WITH AND WITHOUT TIMING

1. Distinction between Verbal and Nonverbal Tests

Verbal tests may be oral or written, and either the oral or the written test may be timed or untimed. The orally-presented test may call for oral or manipulative response from the subject. The performance test may or may not require verbal instructions from the examiner.

There are very few tests in which the use of language is wholly excluded; and such tests as are presented wholly by pantomime can hardly be considered suitable for general use in the clinic, although they may be very useful in examination of deaf children. Without attempting to draw hard and fast lines between the verbal and the nonverbal test, one may yet recognize a distinction between one which is necessarily dependent upon use of language and the one in which language is used only incidentally. The term "verbal test" will be used here to include a task which depends essentially upon the subject's attention to verbal instructions, even when a pencil is used for carrying out the instructions. For example: in a row of five pictures, all clearly recognizable to a child of normal vision, the directions may be to "Draw a line from the cat to the rabbit, going under the bird and over both the cat and the dog." This is considered a verbal task, because it can be performed correctly only by holding in mind the verbal instructions. The same row of pictures could be used in a nonverbal test as one item of a series, the subject being expected to mark in each line the one object that differs essentially from all other objects in the line. In this row of animals the child would be responsible for recognizing the bird as being different from the other animals, but not for stating the distinction between bird and mammal. This nonverbal task involves

fine discrimination in recognition of similarities and differences in familiar objects represented by clear-cut drawings. The task is essentially the same for a page or for a whole booklet; and the requirements can be explained by demonstration, wholly without use of language. When the test is used in an individual examination, however, it is rather unnatural to treat a hearing child as if he were deaf. It is permissible to use language incidentally in the explanation. Thus: "See which of these pictures I mark in this line; now show me which one to mark in the next line." Whether presented verbally or by pantomime, the demonstration should be adequate for a clear understanding of the nature of the task; but after the demonstration the child should be expected to go ahead on his own without further assistance. It is not less a nonverbal test by reason of a verbal introduction, because the task itself does not involve the use of language.

The picture booklet in the **Otis Quick Scoring** series (114) is standardized as a "nonverbal" test, in which the task is to select the different item in each of ninety lines; and is standardized also as a "verbal" test, having specific verbal instructions for each individual line.

2. Significance of Verbal and Nonverbal Tests

The language test is more closely related to school work than is the performance test, and is therefore more important with reference to school placement. For a child of reading age it is highly important to ascertain his reading level as compared with that of other children of the same age. It is equally urgent to find out whether he reads comprehendingly or merely mechanically. For the child who cannot read, the examination should reveal whether he lacks the capacity to learn, or whether he is a case for special effort along lines of remedial reading (p. 5).

For the typical child, the performance test has a special value as a means of catching his attention. If not too difficult to be enjoyed as a toy, the performance task presents a challenge which quickly overcomes the child's self-consciousness; and in many cases it serves to elicit his interest in the examination as a whole. The

results of the manipulative test, however, are not in the same class with the language test findings.

In the early 1920's it was the writer's practice to combine performance ratings from the Pintner-Paterson (121) series with ratings from any independently-standardized language tests that could be found (72). There were then so few tests yielding independent scores that it was not possible otherwise to make up a series from which a "median mental-age" could be derived; but the combination of verbal and non-verbal tests in a single series was abandoned in 1925, as soon as material became available for formulating a verbal battery of seven sub-tests. Observations made in more recent years have strengthened increasingly the impression that manipulative tests measure something different from what is measured by language tests, and that the two sets of results do not belong to the same series.

This writer does not attempt to define intelligence, and does not use the term as applied to anything that can be measured, although recognizing that various aptitudes which admit of measurement are component elements of the complex known as "general intelligence." In terms of our culture, with its current emphasis upon a child's capacity for progress in school, the language test seems to hold the greater possibilities—as compared with the performance test—for some approach to the measurement of what it is most important to measure, called by whatever name.

The writer believes that it is misleading to combine language test findings and performance test findings in a single score, as Wechsler (155) recommends, and prefers that the two sets of findings be kept invariably separate on the ground that the separate scores are more meaningful than the combined score. A comparison of the two sets of results is significant for its bearing upon the subject's aptitudes, interests, and general attitudes. This meaning is lost when a combined score is reported in addition to the two separate scores, because the comparison between two scores cannot compete for attention with a single rating expressed in one figure. A Stanford-Binet report includes the child's age and his mental-age, but the figure that catches attention is the IQ. Similarly in the Wechsler report, the full scale IQ is the score that receives most notice and

is most likely to be remembered. When there is a marked disparity between the two scores, the combined score is misleading. It is as if a child, given the task of adding the fractions 5/13 and 7/25 should offer the answer 12. Verbal achievement and performance achievement are fractions of mental capacity, but it does not make sense to add them together unless they can be reduced to a common denominator, psychologically and not merely statistically.

In a series of 164 full-scale Wechsler examinations made at Danvers State Hospital, the average verbal IQ was five points higher than the performance IQ and three points higher than the average full-scale IQ. In one instance the performance IQ was forty points below the verbal IQ. The older subjects were especially weak in the performance scale as compared with younger subjects.

This was a by-product of an unpublished study of a comparison between the Wechsler-verbal IQ and the Stanford-Binet IQ, 1916 edition. At first the Wechsler scale was used in full. Discrepancies observed between the verbal and performance scales led to a preliminary tabulation, as a result of which all use of the performance scale was discontinued. The writer believed, and still believes, that the verbal scale is on the whole a stronger test than the full scale; but concedes that the verbal IQ—when not lowered by the performance scale—may quite possibly be a few points too high.

The series of Wechsler-verbal plus Stanford-Binet examinations was carried up to 500 cases, all examinations being made by assistants in the department of psychology. Condensed records containing the individual IQ's are still in the possession of the writer. By the time the series was completed the 1916 Stanford-Binet scale had almost passed out of use. The study was of local significance, but there was no urgent reason for bringing it to publication. The findings may be summarized here, for whatever interest they may still hold:

The IQ's were averaged for each age-group according to Wechsler's groupings. Nine groups were included, from ages 15-19 to ages 55-59. There were only 28 cases in the upper group, and 43 cases in the age group 50-54. For the other seven groups the number of cases ranged from 50 to 76. Stanford-Binet IQ's were taken on the 16-year base and also on the 14-year base. The graphic representation showed three curves running from the teen-age group to the near-60 group, with near-parallel fluctuations but no crossings.

The most significant feature of the study is that Wechsler-verbal IQ's showed a higher average for each age-group than the Stanford-Binet IQ

on the 14-year base. As would be expected, the difference between these two curves widened increasingly at the upper age levels.

No clinical examination should be considered complete without both verbal and nonverbal tests. There are subjects for whom a complete examination is impossible, but in such cases the partial examination should be recognized as being incomplete. For the deaf child who cannot read and who has not acquired lip-reading, the language test is useless, and the examination will of necessity be based upon manipulative tasks. Similarly, the child whose hands are incapacitated can be examined only by the orally-presented test. The large majority of clinical examinations, however, should include tests of both types.

There are cases in which the examination has to be improvised. In an institution for defectives the writer once had occasion to examine a boy who was crippled by spastic paralysis. He could walk, with great difficulty, but had no power of speech and almost no use of his hands. He tried to respond to direct questions by head movements, but his attempted affirmatives were not distinguishable from negatives. It was by sending him to different parts of the room that examiner was able to prove his ability to understand verbal instructions. When a simple formboard was presented, he pointed with his elbow in the general direction of the place for a given block.

Although his mental level could not be ascertained, the boy was given an experimental trial in the classroom where children of low mentality were being taught to read a few words from a large chart. A month later some of the words on the chart were printed on flash-cards, and these were spread before him on a table, arranged by chance. He was able in several instances to point out the word named by examiner, thus proving that he was making progress in learning to read.

His place in the school hinged upon this very crude examination. Individual instruction was for children who could be trained for self-support, not for one who would have to spend his life in an institution; but if this boy could learn to read by listening to the instruction given to the group, it seemed abundantly worth while to let him have that opportunity. In view of what was at stake for his mental development and his happiness, the examination which produced no score was of unusual significance.

3. Oral and Written Tests

A language test that carries a single score should be either oral or written, not mixed. A test containing reading matter is useless for the subject who cannot read, or for a literate person who happens to be without his glasses; and a few written items at the upper end of a composite scale may invalidate the score for some particular subject. Similarly, the oral question that calls for an involved response is useless for the subject of defective speech whose response cannot be understood by the examiner; and its usefulness is limited for the subject who is too timid to respond except in monosyllables.

When possible the examination should include both oral and written sub-tests, in order to be sure of using the form that elicits the better response. The examiner can learn something about the subject from observing his reactions to both the oral and the written test. As school tasks for children of reading age include both oral and written work, so the clinical examination should give the child an opportunity to show what he can do by either method.

Many years ago two boys examined the same day, of about the same age and inseparable chums, showed very wide divergence in their reactions to an oral test and a written test which were being tried out with reference to their suitability for standardization. As the tests carried no norms, the scores were evaluated in percentage-of-correctness. One boy achieved 54 per cent in the oral test and 81 per cent in the written one; whereas the other boy's record was 75 per cent in the oral test and 22 per cent in the written one. The difference in their mental levels, as indicated by exhaustive psychometric examination and also by their school work, was not wide enough to be significant; but if the examination had been all-oral or all-written, the results would have been grossly unfair to one of them.

The written test does not necessarily involve an exercise in penmanship for the child. It is wasteful of time to require actual writing on the part of a child who writes laboriously, and there is danger that his record may be illegible. For children below eighth grade the most convenient form of written test is the printed page which calls for simple markings by way of response, not for written words.

Such a test may be used for a child who reads with fair facility, even if he cannot write at all.

4. Timed and Untimed Tests

The ideal examination would include language tests both timed and untimed, and also nonverbal tests both timed and untimed.

Speed is a factor not to be discounted in any type of mental activity. Even in the untimed test which is presented orally, it makes a difference whether the subject responds promptly or hesitatingly. Some of the tasks of an examination should be timed, but not all. In many instances the test scored by speed measures little except speed; and although speed of itself is worthy of being measured, it is not the only thing to be measured.

The effect of a time limit upon a test intended to measure something other than speed may be illustrated by a hypothetical situation:

A "multiple-choice" vocabulary test of 100 items has in each line a key word, followed by five words from which the subject is to select his response. There is no penalization for errors.[1] This test is presented to a high-school group, with a time limit of three minutes.

Subject A covers 97 items, scoring 60 correct responses and 37 errors. Subject B has a score of 50, with some omissions but no errors. Nothing important is learned about these students, because the teacher must have already known one to be a rapid worker and the other a very slow worker. The records do not reveal which of them has the better vocabulary. If A's responses that were correct by chance could be identified and deducted, his score might suffer a loss of several points. If B had been a little less deliberate and more venturesome, he might have earned a higher score. It is plain that A should be instructed to work more carefully and that B should be taught the value of speed. The records may be used to some purpose for their instruction; but in order to ascertain how many of these 100 words either of them really knows, the test should be given without the time limit.

[1] Penalization, if any, would have to be arbitrary. It is more customary to penalize errors in tests of the "true-false" type than in tests having five possible responses.

The use of speed in scoring tests for preschool children is unfortunate, because young children have a very poor sense of time. It is quite an achievement to induce the preschool child to perform a task at all, without regard to the time occupied by his performance.

There are subjects at any age level who cannot do their best work under a time limit. There are children who are seriously disconcerted by the sight of the stop watch, who lose their poise when they have any reason to feel hurried. It is for such children that the untimed test is especially needed. Again, there are children who like to work rapidly and who respond pleasantly to the stimulation of being timed. It is because clinical subjects differ so widely in their tastes and in their reactions that no single test or scale can serve appropriately for all examinations.

The reputed strength of the Stanford-Binet scale for children and the Wechsler scale for adults may be due in no small measure to the fact that time limits have been used sparingly in their standardization. The best of oral scales could be spoiled for clinical use by setting a time limit for every response. Also, many school tests, standardized without time limit, would make excellent instruments for clinical examinations.

Even when presented as a group test, the written battery having no time limit may produce results comparable with those obtained in individual examination. In 1927 the writer paid monthly visits to a state industrial school for examination of recently-admitted boys. Those who had entered the school since the last visit were assembled in a schoolroom for a group examination that was to be preliminary to a more intensive individual examination. The boys accepted the group examination as naturally as if it had been a routine school exercise, probably not recognizing it as a mental test. Their reaction to the individual interview, however, was very different and highly unfavorable. Almost without exception they resented the examination as a personal affront, and it was only by means of performance tests that the examiner was able to obtain their cooperation. The Stanford-Binet scale elicited such resistance that it was soon dropped from the schedule as useless, except for nonreaders who had been excused from the written examination. A verbal test was urgent, with reference to placement in school; but it would have been impossible to obtain trustworthy results without this series of untimed tests which admitted of group presentation (73). The situation was of

course exceptional. It was probably because these boys had recently been committed to the school that they were unduly suspicious.

In order to have a means of comparing a subject's speed of performance with his attention to accuracy, the writer has standardized a battery of written tests both with and without time limits (*77*). These tests were group-standardized, in grades three to eight, although intended for use in individual examinations. They were presented first with a time limit of two minutes for each of seven pages, after which the children were given red pencils and were permitted to take their own time for finishing each page. Corrections made with the red pencil were encouraged; but there was no use of the eraser nor of the lead pencil except during the timed interval. The untimed score included all the correct responses on the page, and the timed score was based wholly upon the markings with lead pencil. Independent norms were developed for the timed and untimed scores. The child who completed a task within the time limit achieved a higher mental-age rating by the timed work, although the two scores were the same.

A task thus standardized for clinical use should not be so long as to be too fatiguing to the younger children, nor so short as to permit serious inaccuracy in the timing. The children in the lower grades cannot be trusted to observe strictly the signals for starting and stopping, even under the watchful eyes of several assistant examiners. For a task having a time limit of less than one minute, or for any test presented with item-by-item timing, the error is too serious to be disregarded. After longer and shorter time intervals had been tried, the two-minute interval was adopted as being the most favorable—long enough to reduce the error in timing to a negligible factor, but permitting a breathing-spell often enough to prevent obvious indications of fatigue. A task of from three to five minutes could be used at the high school level, but it is not desirable at any age to keep the subjects working at maximal speed for longer periods than five minutes.

It is as untimed tests that these battery sub-tests are most useful and most used. Timed tasks that have been standardized on a much larger scale can easily be found, but there is a great scarcity of tests standardized without time limit.

5. Distinctive Features of Performance Tests

The performance test involving manipulation of tri-dimensional materials may conveniently be considered as a timed task, because nearly all such tests are timed. All tests referred to in this section are scored by speed if scored at all.

The manipulative task seems to differ importantly if not essentially from the verbal task; and also the various manipulative tasks seem to differ more widely among themselves than do the various verbal tasks.[1]

Performance tests appear to measure aptitudes developed by whatever manipulative experience the subject has had rather than his potential but undeveloped aptitudes. More physicians and medical students have been available to the writer as subjects than any other class of professional men; and it has been observed repeatedly, although the records have not been preserved, that surgeons are strong in these tests whereas laboratory men are notably weak. Cabinet makers of very little education rank high, almost on a level with mechanical engineers; but a man ranking high in the academic or legal field may give so poor a performance as to make himself ridiculous. A student of world reputation once spent three minutes fitting five rectangular blocks into the frame; and a few days later the task was performed in half a minute by a six-year-old girl. The problem could not have been intrinsically difficult, if it could be solved so easily by a small child; but the professor made hard work of it because it was not along his line. In general, women are weak in performance tests as compared with men. However, a dressmaker was once observed to break all previous records in a formboard task which involved very fine grasp of diagonal matchings. This test was directly in line with her experience in laying patterns on goods in different ways to find the most economical plan for cutting the goods.

Usually, and perhaps inevitably, the manipulative test involves some play of chance. For this reason, the single task is a weak test as compared with a series of increasing difficulty, having tasks

[1] This is a personal impression, for which no exact data can be offered in support.

somewhat similar in kind. The early tasks of the series help the subject to become familiar with the nature of the problems and may enable him to perform a task that would be beyond his reach if it were presented first. The series offers some indication of the subject's ability to take advantage of the experience he has had in the earlier tasks. Some part of what he has learned about the problem may have been acquired by chance; but the use he makes of his experience, in the solution of subsequent tasks, is significant for what it reveals of his capacity for acquiring a new skill.

Some performance series are graduated in difficulty from the six- or eight-year mental level up to the adult level. When presenting the series to an adult from whom superior achievement may be expected, it is necessary to start at the beginning of the series. The first two or three tasks may be the merest child's play, but the tasks at the upper end of the scale cannot be scored unless the subject has had all the practice to which he is fairly entitled.

It happens all too frequently, in the standardizing laboratory as well as in the clinic, that there is time for only one full-length graded series. Not many persons of any mental level have acted as subject for the writer in more than one series, and there has been no attempt to establish correlations between any two of the performance tests under consideration here. In a few cases wide discrepancies have been noted, but not quite so wide as the discrepancies observed occasionally between a subject's performance record and his verbal achievement. Casual observation of clinical subjects seems to indicate that the formboard series and the color-cube series measure something different in aptitudes or experience. At the lower end of the scale, the formboard test seems to draw upon experience that is almost universal at the six-year mental level; but the aptitude or experience required for high achievement at the upper level is something that an intelligent adult subject may or may not possess.

The performance test has a value for the clinic quite apart from what it purports to measure. Interest in puzzles, although not universal at any age or mental level, is a very common trait. A task of difficulty suited to the subject—within his reach but requiring concentrated effort—frequently presents a challenge of compelling appeal. Enough subjects respond to this appeal to justify giving the

performance test a tentative trial in almost any case. The subject's reaction to a series of increasing difficulty is highly revealing. A boy of ten years, slightly retarded and badly spoiled, cried like a baby because his mother was not permitted to accompany him to the examining office. Several simple performance tasks were presented in quick succession, only to bring the response: "That's baby stuff, I don't want it." He was finally prevailed upon to perform the first task of the formboard series, on the promise that he would then be given something more interesting. He complied savagely, jamming each block into its place noisily and vindictively. Although he approached the second task in the same spirit, its unexpected difficulty caught his attention. The third task presented a real challenge, requiring sufficient effort to command his full cooperation. His initial unwillingness to submit to the examination was matched by his reluctance to abandon a problem he had undertaken to solve. All signs of irritability disappeared, and his persistence contrasted sharply with the immaturity of his earlier reaction. It was necessary to give him a little unobtrusive assistance on the last task presented, and his achievement on the series was low for his age; but the change in his attitude was far more significant than any numerical results obtained from him. Children of his mental level usually begin to show interest before reaching the third task, even if somewhat hostile at first. Except in degree, however, the case is typical rather than exceptional.

6. Significance of Speed in Performance Tests

The time record has a significance for performance tests over and above that for the verbal test. Manual performance of any type is thought of in terms of speed, possibly because of the importance of speed in the industries. The manual task in factory, farm, or home is evaluated largely according to speed; and the slow worker cannot compete with the rapid worker except by contributing something extra in the quality of product or service. Even when a task is primarily mental and only incidentally manual, it is difficult to dissociate it from speed.

It is widely recognized that the performance test involves more than speed, but the other factors are less easily scored. In our quest

for numerical findings, we are tempted to overlook factors that cannot be evaluated numerically.

A favorite way of supplementing the speed score is to count the moves made by the subject in reaching the goal, or to count the false moves which may be designated as errors. There is no doubt about the significance of the moves and errors, but there is room for question concerning the accuracy of the count. The subject does not as a rule make his moves with reference to the examiner's ease of counting them. On the other hand, he may work with both hands so rapidly that there is an overlap between moves. The examiner's attention is strained to the utmost in following the moves and recording them, and he has no time for taking notes on any other phase of the subject's performance. Anything relating to the subject's mode of procedure is left unrecorded, because of the examiner's effort to obtain an additional numerical score. It would be highly instructive, as a laboratory study, to take a cinematographic record which could be evaluated at leisure; but such a study would not furnish norms that could be used in clinical examinations.

There is one performance test, **Dearborn Form Board #3** (33), that was constructed specifically with reference to an accurate move count. This test will be described in a later chapter. The writer has never devised a performance test that could be scored with reasonable accuracy by any variable other than speed and is not familiar with any such test except the Dearborn board (87).

Many attempts have been made to reduce the time record and the move count to a single score, either by adding a few seconds to the time score for each unnecessary move or by increasing the move count in penalization for an unduly long time spent upon the task. An elaborate but highly arbitrary system of scoring was developed by Carl for the **Hollow Square** test (24), in which both speed and moves were combined in a single score. In a review of this test prepared for the Buros Third Mental Measurements Yearbook (23), the writer made the following statement:

"The fundamental objection to the move count, except in a test designed specifically with reference to an accurate move count, is that it is too subjective to be trustworthy. The movements which a subject may make are too varied in kind to be accurately defined. There is no oppor-

tunity for deliberation concerning a hair-splitting distinction, and no check on a move that has been recorded or left unrecorded. While the examiner is trying to decide what to do with some particular half-move, the subject continues to make more and more moves. It is true that the highly practised examiner may be able to establish his own criteria for evaluating the moves, but it does not follow that two experienced examiners would agree as to their criteria.

"The subject who tries to work rapidly is likely to make unnecessary moves, and the one who concentrates upon economy of moves is almost sure to take more time than is needed. There is an irreconcilable interference between the two variables. Scoring them both in terms of one or the other is an attempt to get something for nothing, and is almost in the class with the perpetual-motion concept. Any formboard which permits an accurate move count should be scored wholly in terms of moves, and the subject should be informed that he is not being timed; whereas the test which does not permit an objective move count may better be evaluated wholly in terms of speed." [1]

If any test can be evaluated both by speed and by errors, the maze would seem to offer some promise for such treatment. If two records were taken, one for speed and one for errors, it would be unnecessary to reduce them to a single score. They should be treated rather as two independent tests, although stemming from a single performance. Using them for separate age-curves and separate norms would double the value of the performance by furnishing two measuring units for the performance battery instead of only one. Why sacrifice two honest scores to obtain one artificial score? Each of the scores is worthy of being preserved in its own right. Pintner and Paterson (121) used the time record and the error count for separate scores, and gave each score a place in the series from which the median rating was derived. This method of evaluation has an advantage even for the examiner who is satisfied with the accuracy of the move count and who accepts the principle of combining the two scores: it is that the speed record remains undamaged on those frequent occasions when the move count is lost.

There is much to be learned about a subject from observing his performance, and much to be reported that cannot possibly be stated in figures. The speed score is of value, but it never tells the whole story.

[1] Quoted from the review, by permission of Oscar K. Buros.

It must be remembered also that the time record may be misleading as an index to the child's capacity for comprehension of the problem. Some subjects cannot fairly be held responsible for rapid work. The speed-scored performance test discriminates against the child of poor coordination as surely as does the orally-presented scale discriminate against the child of defective speech. In any clinic which exists largely for the study of handicapped children, defects of all types are commonly met. In many cases it is a grave injustice to evaluate a child's achievement in a test according to normal standards.

7. Diagrammatic Classification of Tests

- **Verbal tests**
  - *Primarily for nonreaders*
    - Oral presentation, with questions calling for oral response (1)
    - Pictures, drawings, or mazes, to be marked according to oral instructions (2)
  - *For literate subjects*
    - Printed page requiring some writing on part of subject; arithmetical problems, supplying deleted words, etc. (3)
    - Multiple choice tasks, the significant word in each item to be marked but not written by subject; word or sentence building (4)
- **Nonverbal tests, for literate or illiterate subjects**
  - Manipulative test, having tri-dimensional apparatus to be handled by subject; the task, as opposed to the materials, may be in either two or three planes; arrangement of pictures according to sequence, if on separate cards (5)
  - Pencil-performed task presented without essential verbal instructions, with no reading matter; maze presented by demonstration, copying designs, supplying missing parts of pictures, recognition of similarities and differences in pictures or drawings (6)

This chart is offered for the aid of examiners in selecting sub-tests for a battery. In any category a test may be timed or untimed. Actually, however, those of the first group are usually untimed; and those of the fifth group almost invariably timed. Within each group, it still remains to find a test not too far removed from the subject's mental level. This requires a little trial-and-error procedure on the examiner's part in almost every case, and there should be several presumably-appropriate tasks within reach. When a task elicits an unfavorable reaction, the examiner should be prepared to shift quickly to something more acceptable. It is helpful to keep at hand a folder containing one-page printed forms both verbal and nonverbal.

The equipment of the clinic should include something from each of the groups described in the chart, with a varied stock of printed materials. Some of the many group-standardized tests intended for use in schools are serviceable in the clinic, but some care is required in selecting them. The single-page test is more likely to be useful than the booklet of several pages. As new tests become available, it is worth while to obtain sample sets.

Still greater care is needed for equipping the clinic with high-priced apparatus for manipulative tests. Materials having a wide coverage of discriminative capacity are to be preferred in general to those intended for a narrowly-limited age range. However, the clinic specializing in the study of young children will need a good assortment of tests standardized at preschool levels, although these tests may be useless for older children. The cost of any instrument should be balanced against the frequency with which it is likely to be used.

CHAPTER VI

USE OF UNSTANDARDIZED TESTS

1. "Shock Absorber" and "Buffer" Tasks

There will always be a place for the nonstandard task at the beginning of an interview, even when a standard composite scale is used for the examination. It is unfair to the subject to include in his score a task that is presented before he has had time to adapt himself to the situation, and therefore the "shock absorber" task should be considered an essential feature. The "buffer" task at the end of the examination is equally important when needed, although not invariably required. The final task of the formal series may stand as final when the response is obviously successful; but a failure at the close of the scored series should always be smoothed over for the subject by a task selected wholly on the strength of its appeal for him. By this time the examiner should know enough about the subject's aptitudes and tastes to enable him to make a good choice of a final task.

The preliminary task may present unexpected difficulties. The child may have more reason for feeling fearful and uneasy than the examiner suspects. It may be that his visit to the clinic has been imposed as a punishment, possibly after having been held over him as a threat for many weeks. Frequently the case has been badly mishandled; consequently the examiner has to bear the brunt of mistakes made by others. It is always a safe guess that the child has been brought to the clinic against his own wish and that some diplomatic effort will be required for winning his confidence and cooperation.

As an aid to the examiner in overcoming an attitude of hostility, there should be at hand a varied assortment of attractive tasks, both verbal and nonverbal, to be used wholly without reference to any fixed rules for presentation. Not until something has

been found which elicits a favorable response should the child be considered ready for the tasks that are to be used in his score.

For the subject of marginal cooperation, the informal task is useful at critical points all through the examination. As often as the subject appears to be on the verge of refusing further cooperation, it is advisable to introduce something less exacting than the regular tasks of the scored series. The unimportant task is restful to the subject, although he knows not why.

2. Advantages of the Nonstandard Task, for Informal Use

Standardized tasks that can be used in a flexible examination are not so abundant that one can afford to use them wastefully. There is no mental level for which the supply of sub-tests is more than adequate. All available tasks that have been independently standardized are needed for the battery from which the score is to be derived. Nonstandard tasks would be important for supplementary use, even if they possessed no inherent advantages. There are, however, more urgent reasons for having unstandardized tests as a part of the clinic equipment:

The use of standard tests involves a point of honor. In a flexible examination the examiner may select those sub-tests that he knows or guesses to be suitable for the subject, but he may not play fast and loose with the scores. Having presented a given task, he has committed himself to including its score in his report. To discard the score of the sub-test in a flexible battery is as irregular as to leave unrecorded a failure in a Binet examination and substitute an alternate for the item.

To be sure, it is more important to be fair to the subject than to be fair to the examination, in the event of an absolute conflict between the two ends. The subject should not be held responsible for the examiner's mistake in the selection of some particular sub-test. The examiner should rather hold himself responsible for finding out, before presenting scored tasks, what tests will be fairly discriminative for the subject. If the conflict can be prevented, it need not be evaded. Errors in selection of test material can usually be avoided by an adequate supply of unstandardized tasks, carefully used. In

this way the examiner can be fair to the subject while keeping faith with himself.

The freedom with which a nonstandard test can be used is an asset to the examination not easily overrated. Every standardized scale or sub-test imposes rules and restrictions which no examiner is at liberty to violate; whereas in the use of unstandardized material any examiner may be guided by his own experience and his own common sense. He may not need to take advantage of it in what may be considered a "normal" examination, in which the child gives good cooperation as a matter of course and in which it would seem a waste of time to make much use of nonstandard tests. However, it may be a matter of crucial importance in the occasional examination which presents unusual difficulties, and in which the subject is abnormally suspicious and unresponsive. Such cases are exceptional, but they are usually the cases in which an adequate and trustworthy examination is most urgently needed. So important a matter as whether a youthful delinquent should be placed in a foster home, or whether he should be committed to one of two state institutions, may hinge upon his capacity for responding to training as indicated by psychometric findings. The realization that something serious is at stake for him will be more likely to anger him than to stimulate him to put forth his best effort. The examination will not offer a fair index to his mental capacity unless the examiner can manage to arouse his interest in the tasks, so that his resentment may be forgotten for the moment.

The examiner, not less than the subject, needs the relaxation of the task which is not an essential part of the examination. The subject is not informed that a given task is not to be scored, because it might irritate him to have the interview prolonged by an extra and unnecessary task. The subject feels relaxed because he finds the task easier and more attractive than most of the examination; and the examiner feels relaxed because he is free to adapt the presentation to the subject, as he cannot do in presenting a scored task. The tension of the situation is appreciably relieved. Without interrupting the continuity of the examination, the informal task has the effect of a rest period. The subject has better courage for the next task.

It is for this easing of tension—at the beginning of the examination, in the middle or at the close—that the nonstandard task has its greatest value. It also serves another purpose in its effect upon examiners. It is an important part of a psychometrist's training to form the habit of observing closely what the subject does, and to write reports based upon observation rather than scores. When using a standard test the examiner is concerned with taking an accurate record, and may not have much time for supplementary notes. Heavy routine work in presenting a standard scale has a tendency to make the examiner somewhat indolent and perfunctory in his observations, especially when the report is intended for someone whose interest is centered upon the numerical findings. On the other hand, a performance that cannot be reported in terms of norms forces the examiner to give close attention to the subject's reactions.

As a method of training student-assistants to observe, the writer used to require them occasionally to present without the stop watch a formboard series ordinarily scored by speed. Their helpless bewilderment showed plainly that they had been allowed to become excessively dependent upon the stop watch in taking their records. When deprived of their regular scoring device, they thought at first that there was nothing to be recorded or reported. Eventually, however, some of them learned to write excellent descriptive reports. Many times a marked change has been observed in an examiner's attitude toward the examination, as a result of the request to report on a performance that could not be scored.

3. Effects of Standardization upon Test Development

Every test has its beginning as an unstandardized test. In the early stages of its development a test is being tried out on new subjects and is being revised constantly in the light of the findings. Items that elicit satisfactory responses are retained for standardization, and any items to which the subjects react unfavorably are weeded out. The test is in a state of change and is being improved with continued use. Very few tests, however, are allowed sufficient time for this normal growth. There is usually an urge to standardize them prematurely; and no further revision is possible after the col-

lection of data for the norms. The tabulation of the records taken for standardization may bring to light many defects that are recognized as such only after it is too late to correct them.

The ideal clinical test, like most ideals, is something to be approached rather than attained. Here, as elsewhere, the ideal serves its purpose if it keeps us ever struggling toward an ever-receding goal. But the standardization of a test is conducive to stagnation rather than to progress. Standardization on a large scale is a rare privilege; and the many who use tests are expected to use such instruments as have been furnished them by the privileged few. In child-guidance clinics and in all institutions for children there is heavy pressure for numerical results rather than for an individual report that might be of real service in understanding the child. To an unknown extent, clinicians are using tests that have little to recommend them except the standardization, and are letting their own inventive resources go to waste. So little notice is taken of an unstandardized instrument that it is hardly worth while to publish it, whereas a standard test may continue to be used even after it has been declared obsolete by its author. Statisticians who have access to the data are opposed to a change in the method of examination. From their point of view, it is more important to preserve the continuity of a long-range statistical series than to give the children the advantage of an up-to-date examination. In one known instance, the scale adopted in 1920 for state-wide use held its place as the "accepted standard" for a full generation. One can recognize that the standard test has a value, and yet may question whether the norms are worth what is being paid for them.

As noted previously (page 38), language tests may pass out of date for reasons not foreseen. The danger of this increases with the increasing age of the subjects for whom the test is intended. It is only at preschool levels that either the questions or the norms stand a good chance of retaining their value for long periods. The information ordinarily acquired by very young children is fairly uniform from one decade to another. Individual scores differ widely because of differences in environment, and therefore no great constancy of results within a group can be expected at any time; but the passage of time as such is not so disturbing to norms for

preschool as for school-age children. It is when children are old enough to read newspapers that the temporal factor interferes seriously with the stability of the norms. It would hardly be possible to revise or restandardize a scale often enough to keep up with the kaleidoscopic changes of our times; but—to revert to Chapter IV—it would be a very simple matter to publish norms for independent sub-tests, short and relatively homogeneous. The out-of-date questions could then be discarded, instead of being permitted to vitiate the results of the whole examination. The flexible battery can be kept up to date, as the composite scale cannot.

Manipulative tests are not so susceptible to temporal changes as are language tests. Any test involving recognition of spatial relations depends for its solution upon individual experience rather than social experience. Such a test may hold its value up to the time when it is displaced by something better. It is not because of age but rather because of inherent weakness that so many of the performance tests used a quarter-century ago have passed into the discard.[1]

A vast amount of painstaking effort is required to produce a test worthy of being standardized for clinical use. Only those who have attempted it can conceive how much preliminary work is needed, how many disappointments lie back of the test that comes to publication, and how often it is discovered too late that a test might better not have been standardized. Norms add greatly to the value of a test that is strong in itself, even while adding to the danger that it will be misused; but when a weak test is standardized, its norms offer potential danger of abuse not balanced even by potential serviceability. To quote from the writer's first paper on formboards, written in 1915 (70): "Standardization, of course, does not render valid a method that is fundamentally defective, but it is all too likely to prolong the life of a criterion which would far better be permitted to die a natural death."

A large-scale standardization project is not a good topic for a

[1] The student who has the phenomenal patience required for perfecting his instrument before standardizing it may count on a long life for his contribution. The Young Maze (169), published in 1922, does not show any indication of depreciation. Although it is only the first task of a series which its author did not live to complete, it is so fundamentally sound that it is not likely to be superseded.

dissertation, nor for any undertaking to be completed within a time limit; but the experience of standardizing a test, even on a very small scale, is of priceless value to the professional clinical worker. One who has had this experience is not likely to assume that any test—even his own—is faultless; nor will he assume that all children of the same age may be expected to achieve approximately the same score. The preliminary work of developing a test is highly instructive to any students who may be permitted to participate. The more flaws can be detected in the test at the outset, the less danger is there that a poor instrument will reach the stage of being standardized. A severely critical attitude on the part of students should be encouraged, let the chips fall where they may. The work of presenting individual tests and of scoring group-test records should be distributed among as many students as possible, both for their own education and for the advantage of having the tests thoroughly checked.

4. Misuse of Standard Tests

All criticism of the current use of standard tests expressed here must be understood as applying only to the clinic. It is not for one of such limited experience in other fields in which tests are employed to criticize the use made of them in education, in vocational studies, or in the industries. This section is concerned with personal observations made in the clinical field since the Binet-Simon scale came into common use. The form first used by the writer was the Goddard edition (47), in the public schools of New York City, in 1915.

If any proper and constructive use has been made of clinical psychometric tests at all comparable with the mischief wrought by them, the writer has had no opportunity to observe it. What has been observed during the past thirty years has been a fetishistic reliance upon numerical ratings, first the mental-age and later the IQ. In a clinic carrying a heavy case load, rapid work is especially at a premium. A "good" psychometrist is one who can examine eight children in a school day.

Here is an illustration of what can happen in a public school clinic conducted by mass-production methods:

An eight-year-old girl, examined by an exceptionally competent psychometrist, achieved a low Stanford-Binet IQ. The report was passed upon by a physician of little experience, who made a diagnosis of mental deficiency. A year later application was made for this child's admission to a school for the deaf, but she was rejected on the strength of that school record. Owing to the efforts of a social worker who had become interested in the family, a further examination was requested. This time a performance test battery was used, and the child achieved a rating slightly above her actual age. The first examiner, failing to notice that the child's hearing was defective, had scored as zero all items that were not answered, strictly according to the rules. The physician also failed to notice the deafness, and accepted the IQ at face value.

Some of the young physicians who assist in conducting the clinics are apparently unaware of the extent to which they rely upon the psychometric findings in making their diagnoses, and perhaps claim that they care very little for the test.

Without defending a situation in which immature workers are permitted to assume duties for which they have not had adequate training, psychologists must accept responsibility for having given them an instrument that is highly susceptible to abuse.

5. Subjective and Objective Criteria

The interpretation of a nonstandard test is frankly recognized as being subjective. The findings obtained by the standardized test are supposed to be objective, but no one knows just what degree of objectivity they actually represent (p. 28-29).

Many students are almost afraid to use or recognize a test that offers no possibility of objective evaluation. They probably assume that if the standard test is susceptible to gross abuse, the nonstandard one must be even more so; that the criterion which has no objective basis at all is more dangerous than that which has some measure of objectivity. This objection to the use of the nonstandard test would have considerable force if it were customary to express the results in figures. Any attempt to derive numerical ratings from an unstandardized test would be the merest guesswork, with no basis even in statistics. However weak the statistical criterion may be in its application to the individual, it is better than nothing—better than nothing, that is, if the results must be expressed numerically.

The nonstandard test would hold almost unlimited possibilities for abuse, if anyone had the audacity to offer a numerical rating based on it; but, to the best of the writer's belief, it isn't done. There is no obvious reason why it should be done. Binet has removed all temptation to use the nonstandard test improperly, and has at the same time greatly reduced any inclination to use it at all. There is more danger that its possibilities will be neglected than that anyone will presume to use it as if it were a standardized test; more danger that the clinical examiner will not be allowed time for using it than that he will be requested to guess at a rating derived from it. For better or worse, the standardized test is here. No test that does not carry norms will be able to compete with it.

If numerical ratings had not been so widely publicized and had not become a fetish, the norms would have inestimable value in the interpretation of individual achievement. They do now have great value for those who are using the rating cautiously as an approximate estimate of the subject's capacity. Norms of some sort are indispensable in the use of a test, in the sense that the achievement of a given child has little meaning except by comparison with what other children of the same age can achieve in the test. The examiner who uses unstandardized tests builds up his subjective norms from his own experience and carries them in his head. The comparison of the clinical subject's achievement with the unwritten norms is a conscious process; but it cannot be expressed in figures and can be expressed in words only with considerable effort.

It is perhaps because of the labor involved in writing a report that psychometrists have so little taste for use of the nonstandard test. In the writer's experience, student-assistants are enthusiastically helpful in the work of standardizing a new test, and they love to use the norms that they have helped to establish; but it is only under pressure that they will write a report on a test that has no norms. Apparently they like to work in ruts rather than in the open field, but willingly do the spade work for the construction of an additional rut.

In view of the overwhelming preference for nominally objective criteria—on the part of both physicians and psychologists, and at all levels of advancement in either field—it seems safe to offer the

mature student all encouragement to cultivate an aptitude for subjective interpretation of test findings. He need not make less use of the standard test but may be expected to use it more cautiously. The supplementary use of nonstandard tests, presented by himself in person, will do more than give him additional tools to work with; more importantly, it should help him to preserve a critical attitude toward any popular brand of pseudo-science.

6. Practical Studies in Group Classification

When children or adolescents are to be classified according to their rank within their own group, an unstandardized test serves the purpose as well as a standardized one. Even when the test selected is one that carries norms, it is well to disregard them. The raw scores usually show finer distinctions than are shown by ratings.

The written battery mentioned above (p. 53) was used for about ten years in classifying the freshman students in a county school of agriculture. The group entering in September included a considerable number of high-school graduates, along with some who had barely completed sixth grade. It was customary to accept any student who could read, down to the nine-year mental level. In order that the students might be assigned to classes and duties in keeping with their capacity, it was urgent to have them classified at the opening of the school year. Absolute ratings were of little importance as compared with their relative rank in the group (81).

The seven-unit battery was presented to the students in the assembly room, first with two-minute timing for each page and then with unlimited time allowance for finishing. Each student was dismissed from the room on completion of the untimed work, and a record was kept of the approximate order in which they finished. The seven scores were totaled, separately for timed and untimed tests. These scores showed a much finer differentiation than the mental-age ratings, which ranged from 9 to 14 years. In the report to the school, each student was given his place in each of the two series of scores. The report on the order in which the students completed the untimed work furnished an additional check on their speed of working.

As would be expected, the students who finished among the first were frequently the ones who ranked high in both the timed and the untimed work. Had there been any occasion to study the records for correlations, they would presumably have shown a moderately high correlation between rapid work and accurate work. It was, however, with reference to individual placement that these tests were given. The examination was important for the exceptional members of the group, rather than for the typical ones. Those of the lower mental levels were to be placed in classes in which the work would be within their comprehension. The very slow workers were to be grouped together in school work, so that they might not hold back the progress of the class; and in outdoor work, to avoid the danger of having their work done for them by rapid workers.

A similar method of evaluating results was followed in the examination of preschool children with reference to their being admitted to the first grade slightly under age (86). The number of applications exceeded the number of seats available for extra children. The test consisted of 25 orally presented questions which required not more than ten minutes for presentation, and the weighted scores yielded a total of 55 points. The scores were arranged in order of rank and reported to the superintendent. The passing score varied slightly from year to year, depending upon the number of vacant seats. The method was very simple as compared with a conventional psychometric examination, but it served well for selecting the children who seemed most likely to be ready to profit by school instruction.

This test was not standardized until later (88), and at the time produced no findings except the score; but if norms had been available, they would not have been used. It was an advantage to be able to tell the parents that this was not an "IQ test," and that it was given primarily to find out what sort of response a strange examiner (or a strange teacher) could get from the child. The test came to be known in the village as the "social-maturity" test; and a child's failure to be accepted was not a serious blow to the pride of the parents.

7. The Test as a Personal Instrument

Every professional worker in the clinical field should have a

few tests of his own, both verbal and nonverbal. Many tests have been developed first for the use of the author, and later standardized. It is more important, however, that the personal instrument meet the needs of the individual examiner than that it be suitable for standardization.

The "Emergency Test" (76) was formulated for personal use and perhaps might better not have been standardized. Inquiries for revised norms are still coming in, showing that the norms are used more heavily than was ever intended. The reply to inquirers might well have been mimeographed as far back as 1936: "Forget about the norms, strike out any questions that you don't like, add questions of your own, use the test as a personal instrument, and keep it up to date by frequent revision. Above all, don't try to squeeze an IQ out of it. Let it be what the name implies, a quick way to ascertain whether a psychometric examination is or is not indicated. You don't need norms for that purpose."

It is not enough that the clinical examination be adapted to the interests and tastes of the subject; it is essential also that the examiner be able to take some satisfaction in it, because he cannot do his best work with instruments that are distasteful to him. Although he may not have time to use many supplementary tests, at least he should have something of his own for the shock-absorber tasks and for an occasional relaxation task.

It is neither probable nor important that any two experienced examiners would agree in all details concerning what constitutes a good instrument. The writer would hardly be willing at gun-point to use the Stanford-Binet vocabulary test in a clinical examination but can name two well-qualified examiners who like this item well enough to use it as the opening task. Doubtless it is a good instrument in their hands, but no instrument can give its best service in the hands of one who dislikes it.

If it seem desirable to eliminate the personal factor and to make the examination as mechanical as possible, it would be better to have each task presented by phonograph records and to let the examiner give his undivided attention to the recording; but so long as the examiner's personality is considered an asset, he should be permitted to conduct some small part of the examination to his own satisfaction.

Having developed an instrument that he likes to use, the examiner may wish to standardize it; but the clinic does not offer facilities for standardization. Before seeking an opportunity for the project, he should have the test used and criticized by other clinical workers. The author is not much better qualified to appraise the value of his own test for general use than to pass upon the real merits of his own children.

The clinic affords an ideal field for the all-important preliminary work of test development. The clinical examiner should be encouraged to invent his instruments and should be allowed time for using them in connection with routine examinations; but—as he values his intellectual honesty—let him be discouraged from developing a test for personal profit. The bias of self-interest is devastating to a student's critical sense. The test intended for profit is not likely to be passed upon freely by critics, but is more likely to be a one-man product. Unless developed as a public service, the standardized test may turn out to be a disservice to the community.

CHAPTER VII

THE PSYCHOLOGICAL REPORT

1. General Plan

The plan recommended for reporting a typical case includes five sections, as follows: introduction, language tests, performance tests, miscellaneous observations, and examiner's impressions.

It is not recommended that a more specific outline than this be adopted, nor that the different workers in a clinic employ a uniform plan. The report should be individual, not stereotyped. To be required to follow an outline imposed from above is deadening to the examiner's interest in preparing his reports; and a perfunctory report written merely as a routine duty is not likely to stimulate much interest in the child on the part of the reader. Any competent examiner can write a better report when permitted some measure of spontaneity than when expected to conform to certain rules that may not fit the case.

There is one rule, however, which every examiner will do well to adopt for himself: that he keep in mind at all times as clear a distinction as he can reasonably make between what he observes and what he merely infers. Of necessity, the factual report will be more or less colored by subjective evaluation; but there should be no avoidable confusion between observation and inference. "Examiner's impressions" may be reserved for the final section.

It is recommended also, for reasons stated above (pp. 47-48), that language tests and performance tests be reported separately, each in a section reserved for that purpose. The fourth section, if needed, is for casual observations made in the course of the examination but not already reported in the sections on findings. With the exception of the first section, which may occasionally constitute the entire report, any section may be omitted when there is nothing to be said

under its heading. Within each section, the arrangement may be informal.

The child's age—not merely his date of birth—should be placed where the reader can easily find it as often as he forgets it; but there should be no special place where a numerical test rating can be found at a glance. As is well known among those who have worked in clinics, some physicians and social workers will seek the rating as the only important feature of the report and will read no farther after having found it. The person who is interested primarily in the numerical rating is not usually the one who can be trusted to make proper use of it. Ratings are dangerously misleading, in the hands of those who accept them at face value; and it is important for anyone having access to the records to know what the examiner thinks about the significance of the ratings.

For the examiner, therefore, the preparation of the report is an opportunity and not merely a routine task. It is his privilege to write a report that will be read with care, one that will draw the reader's attention to the child as a person rather than as a case to be diagnosed. The report should be clear and well organized, but not verbose. It should contain all significant information, but no inconsequential details. The language should be simple and nontechnical, wholly free from diagnostic terminology. The results of mental tests should have a very small place in diagnosis, and it is not for the psychologist to call attention to such diagnostic value as they do actually possess. He should try rather, within his limited field of influence, to undo the harm which has been done by the over-sale of mental tests to the medical public.

The limitations of our resources for examination should occasionally be mentioned in the report. It may safely be asserted that no clinic is equipped to give each and every child a fair and adequate examination, because some of the tests which are greatly needed do not exist. Very frequently the time allowed for the examination does not permit the examiner to take full advantage of the facilities which he has at hand.

It should be considered routine that no single test or single-score scale should be reported as final except when it is impossible to offer anything additional as a check on it. As often as the examiner is

denied the time needed for supplementary tests, he should call attention in his report to the fact that the rating is based on a single score, unsupported by additional tests.

A wide divergence among a subject's ratings should be reported, and should not be concealed by a report based upon a median or average. The width of range has a significance worth reporting, not only with reference to the achievement of any particular subject but also as showing the importance of including an adequate number and variety of tests in any examination. The possibilities of divergence in the achievement of a given individual may thus be brought to the notice of the reader, and will tend to discourage an uncritical acceptance of the findings by any one single test.

2. The Introductory Section

The description of the child as first seen by the examiner is inevitably somewhat subjective as compared with the report on the test findings; but the examiner should try to make it factual, reserving his personal impressions for the section that is frankly acknowledged to be subjective.

The opening paragraph is more than an introduction to the report, because it frequently serves to introduce the child to the physician. As a matter of common courtesy, an introduction should be pleasant. The child should be given every opportunity to make a favorable impression, and there should be no word which might tend to give the reader a trace of bias against him. If the child is disagreeable, the physician may discover it for himself without having it forced on his notice.

In describing a child's appearance and his reaction to the situation, adjectives should be used sparingly except for reporting something favorable. The child may be described as attractive, but not as unattractive. It may be stated that he was alert and interested, but not that he was dull and apathetic. The behavior of a child who makes an unpleasant scene may be described by reporting what he actually said and did, but the child should not be characterized as having an ugly temper or a mean disposition. The behavior is observed, but the disposition can be only inferred. A child can be

superlatively annoying, but the examiner should be careful not to let his report be influenced by his feeling of annoyance. Keeping in mind that he does not know what is back of the child's behavior, he should suspend judgment. When the child is so troublesome that it is difficult to say anything in his favor, the examiner should accept it as a challenge to his own sense of fair play. Let him ask himself what he might possibly have done to elicit such hostility on the part of a child. Inasmuch as the child is not invited to hand in a written report concerning his conduct, let the examiner be careful not to take an unfair advantage of the child. In the event of utter failure to obtain cooperation, the examiner should report it as his own failure, perhaps with the suggestion that another examiner might have better success.

Any special handicap may be mentioned in the introduction, such as defective speech or poor coordination. This is the appropriate place for recording anything in the child's attitude which relates to the examination as a whole rather than to some particular task. The report should contain descriptive matter from which the reader can infer that a child is friendly or pugnacious, confident or diffident, voluble or taciturn, aggressive or shy. There is comparatively little to be said about a child whose behavior is perfectly normal, but any extreme behavior is worthy of mention. The child's adaptation to the situation may be noted here, and his susceptibility to fatigue. When his cooperation is so unsatisfactory as to impair the validity of the results, it is well to state the fact in this section, although the details belong to the sections on findings.

In writing the report, as in conducting the examination, the examiner should be careful not to hold the child responsible for what is beyond his control. It may safely be assumed that serious misconduct is due to faulty training, if not to an organic condition. The child's personal appearance may be unpleasing, but through no carelessness of his own. Rather than describe him as untidy, the examiner may report that he appears neglected or ill cared-for.

As the numerical findings may be misleading with reference to a child's intellectual status, so also may a carefully-written descriptive report convey a false impression concerning his social capacities and general promise.

A boy who was superficially unattractive, tall for his age and poorly coordinated because of a severe birth trauma, was examined first by a psychologist and later by a physician. The psychologist's report opened with the sentence: "Charles is a loutish boy who entered the room with a silly grin on his face."

Thus was the child presented to the physician; and it appears from her report that she was sufficiently suggestible to be influenced by it. There is nothing on record to indicate that either of these women made any attempt to understand the boy. They gave adequate time to their study of the case, and reported at considerable length; but they appear to have regarded Charles merely as a case to be properly and conscientiously classified, rather than as a child for whom there was something to be done.

Within a month of this examination, Charles was passed on to the writer, who can testify that he showed exceptional responsiveness to a sympathetic approach. Although his physique was against him, he had a pleasing personality which had entirely escaped the notice of both examiners in the earlier study. The psychologist could not have reported that he was an attractive boy, but she could honestly have described his facial expression as a "friendly smile" instead of a "silly grin." There was much of interest in this boy, had she taken the trouble to look beneath the surface. Her task was to find it, and to report her findings in such a way as to arouse the physician's interest in the child.

It was for foster home placement that Charles was referred, and he was worthy of a better home than would have been chosen for him on the strength of that unfair report. If this record is typical of her reports, the psychologist has much to answer for. The three unpleasant terms in her opening sentence might have cost the boy all opportunity for anything approaching normal development.

3. Findings by Language Tests

If tests could be used with strict honesty, it might be recommended that only three levels of achievement be recognized for the report: high, average, and low; with no objection to the additional categories "very high" and "very low" for a few exceptional cases. The mental-age rating would be useful for the examiner in classifying the subject; but these three simple categories would furnish all the information needed for the practical serviceability of the report. The qualitative findings would occupy a more important place in the study, and the danger of giving misleading information to persons

not technically familiar with the tests would be greatly reduced. There would be no claim of a finer discriminative capacity for the tests than they actually possess.

The honest use of tests, however, is a utopian concept not likely to be approached in the near future. For the majority of cases in the majority of clinics, it may be assumed that the use of the Stanford-Binet scale will be required and that the IQ will have to be reported. The problem to be considered here is how to make the most of the report.

It is possible at least to avoid the worn-out expression: "The basal was established at five years," and instead to specify what items at the six-year level were scored as failure. The straggling correct responses at the upper end of the subject's range are frequently more significant than the score itself. Many responses both correct and incorrect may be worth quoting in full. Even when a stereotyped report is entirely acceptable, the examiner should try for his own sake to make each report individual and distinctive. Two children of the same age may achieve the same final score, but their responses will usually show differences permitting the examiner to write different reports for the two cases.

The term "scatter" may well be avoided, because of the technical meaning which it has acquired among psychiatrists. It is the kind of scatter that is significant, not the mere fact of scatter. An exceptionally good rote memory for digits will lengthen the range at the upper end, and a very poor rote memory will lengthen it at the lower end. The report of a long-range examination, under whatever name, should be accompanied by an explanation.[1]

Any discrepancy in the child's achievement should be mentioned in the report, for what it may signify: such as a correct reversal of five digits after a failure to repeat five digits forwards. When there is any such indication that a failure must have been due to

[1] This is illustrated by two records which were handed in together for inspection. Two boys of about 16 years were examined in court on the same day by the 1916 Stanford-Binet scale. One of them scored on all the digit items for the entire scale; and the other failed to score on the digit item at the seven-year level and all higher levels. Both records showed a very wide scatter, with meanings as different as possible. The boy who could not remember digits achieved the higher total score.

inattention, it is important to report it. The score, of course, cannot be altered.

The different tests included in the report should be grouped according to the kind of rating they produce rather than according to the test from which they are derived. It is meaningless to report the Stanford-Binet mental-age and IQ in the same sentence, because they are not comparable. When the Wechsler-verbal scale is used as a check on Stanford-Binet, as is recommended for any child having a mental level upwards of ten years, the report may read as follows: "His IQ is —— by Stanford-Binet, and —— by Wechsler-verbal." Thus the two IQ's are placed as close together as the sentence construction permits. The reader will be pretty sure to compare them, and to consider the significance of each in the light of the other. Whether they show a close agreement or a wide difference, each furnishes a valuable check on the other.

All the mental-age ratings by language tests may then be reported in another paragraph. For children who are too young for the Wechsler scale it is of special importance to have several supplementary tests which will serve as a check on the Stanford-Binet mental-age, inasmuch as the IQ by that scale will have to stand alone.

The examiner should be constantly on the alert for qualitative findings which will enliven the report. Everything that brings out the individuality of the child will be helpful to the physicians and social workers in their study of the case. Of course the report should not be so long that no one will have time to read it; but within moderate limits the qualitative report on all language tests should be as rich and informative as the examiner can make it.

For any child of reading age there should be a report on his ability to read; and if his reading level is within normal limits for his age, it is important to ascertain also whether his comprehension of what he reads is equal to his recognition of words. It may not be possible to evaluate his comprehension quantitatively, but it is always possible to include an informal report.

In the clinic in which the IQ report is not required and in which no particular scale is invariably used, the examiner is more concerned with the adequacy of the examination and less anxious about the form of the report. The requirement still holds, however, that

every rating reported should be carefully checked. The examiner is more personally responsible for both the examination and the report than when he is expected to follow a prescribed plan of examination.

The Stanford-Binet scale may occupy a smaller place in the examination. At the early age levels it may be used as an abbreviated scale of four items for each age, in order to allow time for supplementary tests. It should be reported specifically as a "four-point" examination, and the mental-age rating should be included in a series of similar ratings by other tests.

For children upwards of ten years, written tests may be used freely, always with care to use enough of them so that the rating will not be unduly dependent upon any single test. Timed tests should be reported separately from untimed written work. Some of the longer and more adequate tests may be reported singly, in a series with the other ratings which serve as a check on them. Very short written tasks may be grouped into a battery, with a report on both the median rating and the range of ratings. It is desirable that the group of tests thus combined be an odd number, so that the median rating will be expressed in years rather than half-years. The rating by months or half-years is appropriate for young children, but for the adolescent it implies a degree of precision which does not exist. Anything over the year may be indicated by plus.

The form of report for written tasks will be considered further in a later chapter.

4. Findings by Performance Tests

Quantitative ratings are of minor importance in the report on performance tests, except in examination of deaf children. When the report by language tests is adequate and satisfactory, the use of performance tests may be somewhat informal. The recording, however, should be done with all the care required for a formal report. The scores may prove to be of greater significance than can be foreseen. In the event of a marked difference between a subject's ability in language tests and in performance tests, the two sets of scores should be in comparable form. The performance test is a valuable check on the language test, when they are expressed in the same

terms. The comparison between language tests and performance tests may stand at the end of the section on performance findings, with whatever repetition of ratings already reported is necessary to show their relationships.

There is no performance scale for children that is widely accepted as being essential to the examination. Within the limits placed upon him by lack of materials and shortage of time, the examiner is usually free to use the tests which seem to him most suitable and to report the findings in his own way. He will not often be called upon to convert the mental-age into an IQ.

The Wechsler scale is in a class by itself, in that the IQ is derived directly from the score and not by way of the mental-age. When the Wechsler scale is used in full, the verbal IQ and the performance IQ may be reported in the same sentence for comparison; but in a series of performance tests scored in terms of mental-age, the Wechsler performance scale may be included as a mental-age rating.[1]

There is almost no limit to the qualitative findings which may be reported on the subject's mode of procedure in solving a performance problem. What to look for in the performance has already been noted (pp. 19-20). Improvement with practice is highly significant, whether observed by presenting the same task several times, or in a series of tasks having something basic in common. The development of a child's interest, in response to the increasing difficulty of a series, is in many instances the most revealing feature of the entire examination. The formboard series and the color-cube series are excellent instruments for showing a child's persistence—or his lack of persistence—in the face of a difficult situation. Frequently there is so much to report that the examiner is embarrassed by the wealth of material at his disposal and is afraid of making his report too long.

The simple formboard used at preschool levels occasionally enables one to observe a complete learning process within a time inter-

[1] In general the Wechsler performance scale is not recommended for children, but it may be useful occasionally for an adolescent subject. It is too time-consuming for routine use. The Wechsler verbal scale is recommended as a check on the Stanford-Binet IQ, but only for those clinics in which the IQ report is required. Other checks are available for mental-age ratings.

val of a few minutes. The child may start by placing any block over any hole and trying to force it in by pounding with his fist. His first successful placement may be achieved by chance, but he finds it a satisfying experience to feel of the block and know by touch that it is correctly placed. Gradually he shows better comprehension for the task, and begins to use his eyes to advantage. The final performance may show him working with unerring accuracy, taking each block to its own place. Such a performance is worth reporting in detail.

Among the various formboards which stem from Seguin, the Witmer board (167) is the most convenient in respect to size and is by far the best standardized. It is presented for three trials and is scored both for the first trial and for the best of the three trials. The circle is the easiest block to place, and the one which is most frequently jarred into its recess by pounding. The five-pointed star may be resting over its recess at such an angle that a turn of one-eighth inch would let it fall into place, and yet the child may fail to insert it.

Someone has said that the simple formboard is not a test of intelligence but rather a "stupidity test." Called by whatever name, it is a superb instrument for observation. Only a small part of what the examiner learns from the performance can be expressed in a score.

At the upper end of a formboard series increasing in difficulty, one may occasionally observe a mature subject of high mental level making a performance almost as crude as that of the child. The subject who grasps the early problems of the series at sight may run through them so easily that he fails to learn from them the relationships which constitute the basis for the more difficult tasks ahead. Consequently, when confronted by a problem too complicated to be solved without effort, he falls back helplessly on the trial-and-error method.

The color-cube test, which has been used by several students for a standardized series of tasks, ranks almost with the formboard series with regard to the opportunity it offers for objective observations that do not yield numerical findings. The performance is usually scored by speed, but the speed score is less significant than what the examiner observes. One form of procedure that commands the respect of the observer although it is uneconomical of time is

to start the performance by turning each block right side up. This involves handling each cube twice instead of once. It has been found experimentally that a trained subject who is familiar with the designs can perform the task more rapidly by picking up each block only once, turning it right side up and then immediately placing it. And yet, even knowing that the subject is losing time, the examiner likes to observe the systematic performance. It is more satisfying esthetically than is the unplanned performance which may earn a higher score.

The examiner who uses performance tests for what he can learn about the subject will often observe something worth reporting not included in the score.

5. Other Observations and Impressions

The examination offers a fine opportunity for incidental observations, some of which may reveal something of the child's personality. Under any heading implying "miscellaneous observations," the examiner may put on record anything he has learned concerning the child's tastes and interests, his effort and persistence in response to tests of different types, and any spontaneous activity or verbal production not already reported. These observations may be described in brief or in full detail, according to the degree of the examiner's interest. The report on nonstandard tests may be included here.

The chief purpose of the final section, entitled "examiner's impressions," is to give the examiner a place in which to register his dissatisfaction with the examination. When the child gives such good cooperation that the results can be accepted as being truly representative of his ability, there is little to be added. It is when the examiner believes that the numerical findings fall short of the child's potential that it is important to put this belief on record, supported by whatever reasons can be given. It is on no account permissible to offer any conjecture concerning how much better the subject might have been able to do. The examiner may express his opinion that the ratings are too low, but no one has the right to guess how much too low. A conjectural estimate should not be expressed in figures.

Speculative explanations of a child's behavior should be offered sparingly, more for the examiner's own sake than on the child's account. It does the child no harm for the examiner to guess that the child is suffering from "emotional conflicts" or from "lack of security," and of course it may be true; but the reader will not pay much attention to the opinion of an examiner who offers such explanations habitually. The examiner who does not wish to have his impressions discounted as coming from an over-imaginative person should be careful to avoid "blanket" explanations.

CHAPTER VIII

CRITERIA FOR CONSTRUCTION OF CLINICAL TESTS

1. Long-range Plan for a Flexible Scale

Much has been said about the importance of adapting the clinical examination to the individual child. The fact remains that no instrument adequate for a complete examination admits of such adaptation. The best that can be done with facilities now available is to supplement the formal examination with a few tasks sufficiently attractive to the child to be helpful in holding his cooperation.

There is great need for a scale constructed so as to permit the examiner to select in each instance a task that will elicit spontaneous cooperation. We need a large and varied collection of sub-tests, each standardized as an independent unit. Instead of spending an hour on one scale which carries a single score, the examiner should be able to select a dozen five-minute tasks, each chosen as being appropriate for this particular child. The system should be sufficiently comprehensive to furnish suitable material, in abundant quantity, for any possible case. Many short scales for oral presentation are needed, for mental levels from preschool to high school. There should be many one-page written tests, ranging from second-grade reading ability up to high-school levels. Batteries of nonverbal tests should be so varied as to furnish an adequate examination for a child who is totally deaf. At whatever mental level, there should be enough sub-tests to choose from so that there would be no occasion to use a task that might happen to be unfair to the child. The results would yield some twelve or fifteen mental-age ratings; and the demand for a single rating could be met by reporting the median of the series.

The development of such a system is not a project for one mind or one university, but is rather an undertaking for cooperative effort

on a nation-wide base. It should be financed by a nonprofit agency or foundation. It is not proposed as being less costly than the standardization of an elaborate composite scale; but is recommended as being more enduring than any single scale ought to be. If the unit of standardization be the short sub-test rather than the system as a whole, the flexible scale can be kept up to date by internal growth. Any new sub-test that becomes available may be introduced into the system, without disturbing the scale as a whole; and any test that has lost its value may be sloughed off, without being missed. Any particular task may be retained or discarded on its own merits.

Although large-scale standardization is not a project for the individual student, there is much for the clinical worker to do, without waiting for a nation-wide undertaking to be launched. In the long run, the success of the plan will depend upon small contributions made by hundreds of clinical examiners who wish to develop an improved instrument for the study of tomorrow's children. It is recommended that every student who does psychometric work in a clinic keep two notebooks always at hand: one in which to note any question that occurs to him as having potential serviceability in an orally presented scale, and the other for notation of any defect observed in currently used tests. Routine clinical workers are the ones who have the best opportunity to see what is wrong with the tests they are required to use; and their criticisms might be of inestimable value to those who are preparing tests for standardization. The examination of the future should be the product of many minds, the more the better.

A language test intended for standardization should be set up in preliminary form and given a try-out in clinics. Mimeographed forms should be sent out to clinical workers who have no bias in favor of the author and no personal stake in the development of the test. There should be at least one revision after each item has been passed upon by disinterested critics, and perhaps another revision after a second experimental trial. A test formulated in haste, however well standardized, is not likely to meet the needs of the clinic. It is well worth while to take plenty of time for the preliminary work

and to have every detail criticized by a large number of students who are clinically oriented.

The forms prepared for tentative trial should contain many more items than are desired for the final product. Test questions for oral presentation have a very high mortality rate, as may be illustrated from personal experience:

Preliminary forms for "Emergency Test" (76) were printed in 1926, containing 100 items selected out of a much larger number. Before reaching publication in 1932 the test had been reduced to 25 items; and in 1941 it was further reduced to 10 items (83). With the aid of this experience a slightly better record was achieved in the development of the preschool test now known as Scale A of the 1946 "Emergency Series" (88): this scale contains 25 items, selected out of 108 recorded questions, plus an unknown number of which no record was preserved.

The author of a test is fairly besieged with requests for an alternate form to be used in "repeat" examinations; but it is naïve to assume that a student who has produced a test has any material left for a second form. Serviceable questions cannot be conjured up at will. In overcoming some particular defect, one is almost sure to introduce another. Many questions seem promising until they are given a trial in the clinic, but are found wanting when one observes the reactions of the children. Finding enough material for an adequate scale is too heavy a task for an individual or a small group of students. It is only by soliciting contributions from hundreds of teachers and clinical workers that one could collect enough material to permit the elimination of all items recognized as being weak.

The inherent quality of test items should be given first consideration as opposed to the number of items. It was Binet's assumption that the quality could be disregarded, if there were enough items. This position can be defended on theoretical and statistical grounds, but not on practical grounds. The time that can be spent to advantage in an interview with a child is limited in every instance by the child's ability to give good attention, and by the examiner's resources for appealing to the child's interest. If the child's attention can be held for only a small number of tasks, additional items serve only to fatigue him. Of course a test may be

strengthened as a statistical instrument by including many items of inferior quality, but this will not necessarily make it a more serviceable instrument for use in the clinic. Much depends upon the individual child being examined. Some children can give good attention during a long examination and others cannot. The scoring unit should be adapted to the child whose attention can be held only with difficulty and only for a short period, so that there may be something to show for every minute of the examination. Additional tasks of questionable value may be used freely for the child who can tolerate a longer examination, but the weak items should not be included in the same score with the important items.

2. Set-up for the Orally Presented Sub-test

It is generally recognized that the "year-scale" (or "age-scale") is flagrantly wasteful of time. So is any scale or item so presented as to require a specified number of correct responses at the base and of failures at the "ceiling." Those used merely to meet a formal requirement of presentation are outside the range of discriminative capacity for the subject, so that they add nothing to the adequacy of the examination. The time spent on these items could almost certainly be used to better advantage on additional tests not included in the scale.

Furthermore, the time wasted on nondiscriminative questions is very much worse than wasted. These are the items that are most annoying to the subject. The too-easy questions are insulting, except at the very early ages; and the too-difficult questions are humiliating to an extreme degree. These unnecessary items make the examination needlessly distressing to the subject and frequently elicit a reaction that makes him less cooperative for the remainder of the examination. The examiner's task of holding the subject's interest is at least doubled by these useless questions at the upper and lower ends of the subject's range.

All this can be avoided by breaking up a scale of wide coverage into two or more shorter scales. A very simple method of finding the scale most appropriate to the subject's level is to have an overlap of several questions between two adjacent scales. For example:

Thirty questions may be arranged empirically in approximate order of difficulty, and then divided into three equal sections. They may be used for two twenty-item scales, each of which has its own norms. The upper scale includes the twenty items at the more difficult end; and the lower scale consists of the twenty items at the easy end. The middle section of ten items is common to both scales. The examiner may not know in advance which of the two scales is the better one for the subject, but by presenting the middle section first he can find out with unerring accuracy and with no loss of time. If the results leave him in any doubt, it is a safe guess that both scales are within the range of difficulty discriminative for the subject; and that by presenting both end-sections he can obtain two measuring units. On no account, however, should he present a section likely to result in a perfect score or a zero score for the entire section.

It is not for the series of questions that a relatively narrow range of coverage is recommended, but for the unit of standardization. In a long series of questions, the wider the range of discriminative capacity the greater are its possibilities for a number of scales.

A series of whatever coverage can be broken up into short overlapping scales. The number of scales will always be one less than the number of sections, because an extra section is needed for the overlap. Any scale will include two sections of items. It is not essential that all sections be equal as to the number of items, but it is important that the items of a given section be approximately equal in difficulty. There should be an appreciable difference in difficulty between any two adjacent sections. The difficulty of each item should be determined empirically before the series is divided into sections, because the division into sections depends essentially upon the gradation in difficulty. Each item is placed in the section with other items of similar difficulty, without regard to the number of items in any section. So long as the gradation in difficulty between adjacent sections is provided for, the number of scales may be as many as the material permits. If a group of students can collect from fifty to sixty questions of established value, ranging in difficulty from four years to high school levels, they may be used

for five or six scales. Such a series of scales, with an overlap of from seven to ten items between two adjacent scales, would cover the large majority of subjects referred to the clinic and would be of inestimable value.

A continuous series of short scales, each having independent norms, offers the utmost of economy in presentation, not only for clinical use but also for standardization. In the clinic it would never be necessary to use more than two sections, and the examiner may do as seems best about going farther in either direction. If an additional measuring unit can be had by presenting half a scale, he will doubtless wish to take advantage of it. If the upper of the two sections presented shows more than half the responses to be correct, it will presumably seem worth while to go a step higher; or if the lower section presented shows a substantial number of failures, it may seem advisable to go a step lower. In either case, it is for the examiner to decide.

Logical classification of subject matter for the various sub-tests is hardly possible for the student working individually, because he is too hard pressed for material that can be used at all to give much attention to what kinds of items combine well in a scale. If and when the clinical psychologists of the country undertake a cooperative plan for test development, it may be expected that there will be enough material available to permit the grouping in a single scale of the items that properly belong together.

With this possibility in view for the future, it is recommended that the unit for standardization be relatively homogeneous rather than highly composite. The examination as a whole should be as varied as possible; but it should be made so by having a variety of sub-tests rather than by putting widely differing items into a single sub-test. The scores may be expressed in a profile, which will show specific strength and weakness.

The grouping of items with reference to subject matter will be taken up further in a later section, but it is important first to consider the classification of material with reference to the type of response which a question elicits.

At any particular level of difficulty, some questions can be answered in a few words; can be answered promptly if at all; call

for responses that can be evaluated on the spot by the examiner; and can be recorded immediately as a plus, a zero, or a number indicating a weighted value. Also, except at very low levels of difficulty, there are questions that elicit involved responses to be recorded verbatim and evaluated at the examiner's leisure. There is a proper place in the system as a whole for items of both these types; but to include both types in a single scale might spoil two strong tests in order to make one weak instrument.

There are clinical subjects of all ages and mental levels to whom the orally-presented test is a severe ordeal. The adolescent boy examined in court is frequently as ill at ease as a candidate for an advanced degree during his oral examination; and the high-school girl referred by the school because of her failure in algebra may find it almost beyond her strength to whisper: "I don't know." The question that calls for an involved response is useless in such a case, because the subject will give up at the outset rather than attempt a solution of the problem. A scale may be made up wholly of questions that give the subject an opportunity to elaborate his response at length, and may bring very significant responses from a subject who likes to talk; but it is poor economy to introduce even one such question into a scale that might otherwise be serviceable for the extremely diffident subject.

3. Criteria for Items of an Oral Scale[1]

A test should be impersonal, intellectually interesting, but emotionally neutral. Natural phenomena, industries, and travel are favorable topics for test questions. The upper end of a scale should be free from puerile questions and should contain nothing insulting to the dignity of an educated adult.

An orally presented scale should contain no reading matter for the subject. The written test is an important part of the examination of a literate subject, but it should be kept wholly separate from the oral examination. It is for the nonreader that the oral test is of special importance. A single item which requires the subject

[1] A considerable part of this section is taken from an earlier publication (76) and is used with the permission of the publishers, Williams and Wilkins, Baltimore.

to read tends to invalidate the examination for one who cannot read, leaving no language test by which he may receive a fair rating.

The question should be easily grasped, in no instance so long or involved as to require repetition. No heavy demands should be made upon the subject's attention for mere comprehension of the problem. What effort he can be induced to make should be conserved for the solution. He should not be required to listen to the reading of a paragraph or sentence calling for close attention. The question should contain no word that is difficult to enunciate, nor any important word that could cause confusion because of ambiguity in meaning or because of similarity in sound to some other word for which it might be mistaken.

There should be nothing gruesome or unpleasant in any item, nor should any question point a moral even by remote implication.

Except at preschool levels, there should be no question calling for one of two specified responses. A question which offers a fifty-per-cent chance of a correct response by guessing is not in general a useful question, because there is usually no way of knowing whether a correct response is a valid one or a lucky guess. The "dilemma" question, however, is very useful up to the six-year level. The young child shows a very strong tendency to repeat automatically the last word of the question, as was observed by Binet. In order to have any value in a test, the question must be so stated that the last word will be incorrect. The question "Which is larger, a horse or a dog?" is a serviceable item at the five-year level. It brings the incorrect response with amazing frequency; and consequently a correct response indicates resistance to the impulse to repeat unthinkingly the last word of the question. Automatic repetition of the last word is not so commonly observed in children upwards of six years, and accordingly the value of the "dilemma" question is lost. When a child answers correctly the question "Does the sun set earlier or later from the top of a hill?" his response may be either a guess or a reasoned conclusion, and does not deserve credit unless the examiner can satisfy himself that it is better than a guess. Consequently, the question is not a useful one.

A question that invites a guess is at a premium, however, if the chance of a correct guess can be reduced to a negligible factor. Many

questions elicit nothing but frank failure from the subject who cannot answer them correctly. If he can give a response that is satisfactory to himself and which the examiner can tacitly accept as correct, it is all to the good.

In general, the technique of presentation should be simple. Items that are fatiguing either to subject or to examiner should be included in a test only when they are believed to have exceptional value. Most of the questions should be easily presented and easily recorded. Questions which elicit responses difficult to evaluate should be used sparingly, because they are conducive to error in scoring. For the large majority of sub-tests, the examiner should be able to score the responses without referring to a key.

Questions pertaining to nature study are notably satisfactory if used at the right time and place, but are very troublesome at other times and places. On the whole, it is well to avoid items of seasonal or local significance in a test intended for standardization.[1]

At the preschool end of a scale, the question that can be answered in one monosyllabic word is highly desirable; and the word that answers the question correctly should be so distinctive in sound as to be easily understood when spoken by a child having defective or infantile speech.

4. Criteria for Evaluation and Weighting

In the orally-presented scale the individual response should be recognized as the unit for evaluation. There should be no "two of three" or "three of four" requirements for credit, because it is unfair to deny credit for any correct answer. Each response should stand on its own merits, without reference to any similar or related item.

The multiple item is useful in the presentation of a scale, although it should not involve arbitrary withholding of credit for a correct response. Several questions of a given type may be grouped for presentation rather than scattered throughout the scale. There is a

[1] The lower scales of "Emergency Series" (88) were formulated for the use of clinics in Massachusetts and were not intended for publication. Several questions were found to be wholly unfair to children in southern Florida, especially questions involving the names of seasons. Southern children do not learn the names of the seasons so early as do children in the North; nor are Florida children so well acquainted with robins and oak trees.

certain "mental set" which carries over from the first question to the others of the group. It is economical of time to take advantage of this factor in presentation; but the score should be for the individual item and not for the group. Credit for a correct response should not be forfeited because of failure to earn a score on the other items of the group.

It is not necessary, if possible, to equate the values of different items in a scale by statistical weighting. The use of the short subtest as the unit for standardization reduces to a minimum the occasions when a record is spoiled by an accidentally-invalidated item. Should this occur, it is safer to sacrifice the sub-test than to salvage it by the very questionable procedure of pro-rating the remaining items. Any manipulation of results has a tendency to distort the true values of the items; and at best it greatly increases the difficulty and uncertainty of interpretation. The scores of individual items are more meaningful if the responses are recognized as having unequal values.

Weighting of each item according to its difficulty, without reference to the other items in the scale, increases the immediate significance for the examiner of a correct response to any difficult item. Inasmuch as such weighting will in any event be approximate rather than exact, it is advisable to give some consideration to economy and convenience in the work of standardization. For the oral scale which is to be standardized in a small way, the first hundred records will serve passably both for arranging the items in order of difficulty and for establishing their values.[1]

The range of values to be recognized is optional. If four points be the highest value to be used in the scale, the plan employed for the weighting may be as follows:

Questions answered correctly by:
- over 75 per cent, 1 point
- 51-75 per cent, 2 points
- 26-50 per cent, 3 points
- 25 per cent or less, 4 points

[1] The tabulation of the records is a very difficult task for one pair of eyes; but it can be accomplished very easily by two persons, one reading aloud the responses to each item and the other checking off the correct responses on a specially-prepared form.

The convenience of using one hundred records for this purpose is obvious. The forms for the first records should be mimeographed, not printed. The children selected for these preliminary examinations should be near the middle of the range to be covered by the scale, in order to avoid any danger of a perfect score or a zero score. After the first hundred records have been obtained, the preliminary test forms should be discarded. New forms should be prepared, having the questions arranged in approximate order of difficulty and having each item followed by its score value. It is worth while to copy onto the new forms all records to be used in the final tabulation.

Weighting of items is in no way essential, nor is it necessarily recommended for scales of all types. Its principal advantage is that it widens the coverage of a scale. A small number of questions, with weighted values, can be spread over a wider range of mental levels than could be covered by twice as many unweighted items. In view of the scarcity of good questions for the oral scale, there is abundant reason for subjecting the available material to as much stretching as it will bear. Anyone having enough satisfactory questions for a scale of unweighted items can use them to better advantage for two scales having weighted items.

This may or may not hold for a written scale, depending upon the difficulty of formulating the items and also upon the time required for the routine use of the test. It is good economy to use weighted values for a scale that is more than ordinarily time consuming, such as arithmetical reasoning. For a multiple-choice test, on the other hand, it may be easier for the author to formulate additional items to be used unweighted than for the examiner to give attention to weightings when scoring the records. In so far as it is a question of time, the examiner's time should be conserved rather than the author's; but it is more than a question of time when there is a serious shortage of material.

5. Criteria for Arithmetical Scales

Any arithmetical test should be standardized as an independent scale, never as a part of a composite scale.

In the first place, arithmetical ability occupies a place of such importance in the school curriculum and in every-day life as to entitle it to a place of its own in the examination; and the score is more significant when it can be compared with the child's achievement in other elementary school topics than when its identity is lost in a composite score. In the second place, it impairs the essential validity of the examination to include under a single score any task in which it is more-than-ordinarily difficult to obtain the cooperation of the subject.

Refusal of cooperation is more commonly observed in arithmetic than in test questions of any other type. Children of any given age show wide differences in arithmetical achievement. Even in a given school grade—except when ability in arithmetic is recognized as the criterion for promotion—extremes of strength and weakness in this topic will be found. Very low achievement may frequently occur in a child whose school work in other topics is well above the average. The arithmetical test, as presented in the clinic, does not usually reveal whether exceptional weakness is due to incapacity, to defective teaching methods, or to lack of interest. Other children under the same instruction are not available for comparison. Occasionally one can prove that a child's failure is due primarily to lack of confidence and not to incapacity.

The arithmetical test is to the child of notably weak aptitude in arithmetic almost what the written test is to the nonreader. The damage caused by these two items in a composite scale, with reference to the validity of the examination, is perhaps about equal. In degree, the written item is more serious for the occasional nonreader; but the slighter damage caused by the arithmetical item is of more frequent occurrence. In either instance, there is abundant reason for keeping the score separate from all other test scores.

Very rarely should a child be excused from the arithmetical test because of his reaction to it. His unwillingness to attempt it makes it the more important to ascertain his actual capacity. There is therefore an urgent need for several arithmetical scales, both for different levels and also for different types of arithmetical ability.

Most important is the graded series of scales for arithmetical reasoning. In this series the computation should be of negligible

difficulty, so that the measure of a child's reasoning ability may not be vitiated by his inaptitude for figuring. There should be an orally-presented scale discriminative up to nine or ten years and carried down to the lowest possible level. The simplest problem would be to add one and one, in a setting involving real reasoning, as, for example: "Freddy's daddy gave him one penny, and then his big sister gave him one penny." This might be called the game of "How many?" Above nine years the problems may be written instead of being orally presented. The child may be permitted the use of pencil and paper, but the handling of the numbers as such should be as easy as possible. The scale for the age-range nine to fourteen can be group standardized.

Another scale or series of scales for group standardization should be based wholly upon calculation in the fundamental processes. This is a very easy test to devise. The lower end of the scale could be made simple enough for group standardization in second grade, if presented without time limit. For children of fourth to eighth grades the scale should be standardized both with and without timing.

A graded series of tasks in counting money could be made attractive to the child who hates formal arithmetic, and could be used as a means of gaining his cooperation for the more important arithmetical tasks. This would require individual standardization and should be included with other individual tasks. The series could be carried to preschool levels by beginning with three cents, and could be made discriminative at ten or twelve years by a timed task of counting several dollars in small change. There should be two or three overlapping scales, so that the older children need not be required to start at the bottom of the series. The younger children need not be expected to go beyond one error.

For some children, the questions in arithmetical reasoning may be interspersed among orally-presented tasks that are less fatiguing. The series should be scored as a whole, but need not necessarily be presented as a whole.

One commonly used arithmetical task that cannot be recommended is the "number series." There is a tendency among normal children of the lower age levels, and among clinical subjects of any mental level, to fill in the answers to the easy items and then give up.

The test calls for more persistent effort than is often observed in the clinic. It requires both the ability and the willingness to abandon an unsuccessful plan as often as necessary, and to take a fresh start. It is an irritating task, for one who lacks the capacity for making a quick shift from one mental set to another. The test is not serviceable below high-school levels, and is of questionable serviceability for a clinical subject of any level.

6. The Vocabulary Test

Like the arithmetical test, the vocabulary test should have a place of its own in the examination. The equipment of a clinic should include enough vocabulary tests, sufficiently varied, to furnish coverage for all mental levels upwards of two years.

Any attempt to estimate a child's total vocabulary from his responses to a small sampling of words is over-ambitious. If it can be ascertained roughly how well his understanding of the selected words compares with the average achievement among children of his age, this is as much as a single sub-test of a one-hour examination can reasonably be expected to show. Incidentally, and perhaps more significantly, it may show how well his achievement in this task compares with his own achievement in tasks which depend more upon reasoning and less upon his reading habits.

The multiple-choice test, group standardized without time limit, is recommended as a serviceable method of presentation. Vocabulary tests of this type are not lacking for high-school and college levels, but there is a serious shortage of one-page tasks suitable for children below eighth grade.

Vocabulary tests of two levels with overlapping coverage should be standardized in the elementary schools; one series for grades three to six, and the other for grades five to eight. There should be several forms for each level, similar with reference to difficulty but differing in length. The test for grades three to six is most urgently needed, to be used for retarded children up to adolescent ages. The reading matter should be reduced to the minimum, with four words to choose from rather than five. The number of items on the page should range from twenty to forty for different forms; and it is safe

to predict that the twenty-item forms would be more used than the forty-item ones. They should be printed in plain type, with good spacing between words and between lines. The format of the page should be adapted to marginal reading ability. All the words should be carefully selected. The student preparing such a test should have at hand not only the Thorndike word books (148, 149) but also an assortment of reading and spelling books intended for second, third, and fourth grades. At least two of the shorter forms should be based wholly upon words found in the third reader. The easier the task can be made, the more useful it will be.

Development of a multiple-choice picture vocabulary test is an excellent project for the psychologist who is also an artist. For preschool children the task that requires no speech on the subject's part is greatly preferred to the one in which he is expected to name the object pictured, although there is a place in the clinic for tests of both types. Many young children who refuse to speak can be induced to indicate the object named by pointing to it.

The pictures, of course, should be clearly drawn. There may be a row of five pictures on a card, spaced about two inches apart. The objects pictured on a given card should be approximately equal in size, because young children are easily confused by pictures drawn to different scales. At the lower end of the series all the pictures should represent objects with which the young child is familiar.

If a picture vocabulary test can be carried up to higher levels, it should be found useful not only for children who cannot read but also for study of aphasic patients in hospitals. The pictures might show activities rather than simple objects, illustrating words that the young child would not be expected to know; but no picture should present any difficulty of recognition. A vocabulary test is not a test of visual acuity; nor is it a test of the subject's comprehension of obscure relationships. When pictures are used for measuring a subject's understanding of certain words, great care must be taken to avoid putting a premium upon his habitual attention to comic supplements. Aptitude for interpretation of comic pictures, if considered worthy of measurement, should be measured by itself and not confused with a vocabulary test. At whatever level of difficulty, recognition of words is to be measured, not recognition of pic-

tures. The series should be divided into sections and standardized for overlapping scales, with very short scales for preschool levels. Sketchy drawings are not of such interest to a young child as to hold his attention for a long series.

Presentation of words for oral definition is a valuable psychiatric instrument. Many other questions that elicit individual responses constitute an important part of a clinical examination; but it does not follow that they should be included among tasks to be scored, nor that they admit of being evaluated with any claim to objectivity. Frequently a child's responses to the definition test offer more information concerning his self-confidence than concerning his understanding of the words. His willingness to respond is significant, and an attempted definition may be highly revealing; but a failure to respond reveals very little, or may be misleading. When there is no response, the examiner is left in doubt whether the subject is ignorant of the meaning of the word, whether he understands it but cannot think of any other word in which to express his understanding of it, or whether he lacks the confidence to offer a definition which—if he only knew it—would be entirely acceptable. At any mental level there are words that would make sense to a subject if he were reading, although he may not feel prepared to define them. The person who uses the dictionary and who has high standards for definition may be the one who is most reluctant to offer a definition that he knows to be inferior. Even a child of five years may be unwilling to define a chair in terms of use, recognizing that a statement of its use is not an adequate response to the question as presented.

On the assumption that a vocabulary test is intended to show some indication of the subject's understanding of certain selected words, the written multiple-choice test is recommended as offering possibilities for exact scoring far beyond that approached by the orally-presented test requiring individual response.

There is, however, a real need for lists of words selected with special reference to their definability (83). Nouns, adjectives, and verbs should be listed in separate columns, to enable the examiner to do as he pleases about disturbing the subject's "mental set" for words of a given type. Words that are unduly difficult to define may

be eliminated by having several students pass upon them. After each of them has crossed out every word which he recognizes but cannot define offhand to his own satisfaction, the list may be considered ready for informal use in the clinic. It is not recommended that there be any attempt at standardization. The examiner should not decide in advance what words or how many words are to be presented to any given subject. It is a test to be used intelligently, not mechanically.

7. Criteria for Written Tests

The group-standardized written test is needed especially for the child having defective speech, and also for many children who are too ill at ease—in a strange place and with a strange examiner—to respond normally to an orally presented test. The child who reads with fair facility can usually accept the written test almost as naturally as if it were a school exercise; and with his eyes on the page he is less keenly aware of the strangeness of his surroundings. Failures are not so distressing as in the oral examination. The written test is relatively impersonal, as compared with the oral one, and is therefore more comfortable.

In the formulation of written tests the range of discriminative capacity is of prime importance. Although the subject does not react quite so strongly to an inappropriate item in the written test as in the oral test, it still remains that the time allowed for the examination is too precious to be wasted. There is still abundant reason for keeping the nondiscriminative matter at a minimum.

Some written tests intended for group study are planned specifically to cover all mental levels from the earliest reading age up to full maturity. A test that is unlikely to result in either a perfect score or a zero score may be useful for a rough survey of large groups; but for almost any individual it contains useless items at each end of the scale. The written page has to be presented as a whole, and it is not possible to have a middle section common to two overlapping scales. The ideal test for the individual subject is one having a range just wide enough for coverage.

A battery discriminative for the seven-year stretch from eight to fourteen years may be expected to see daily use in a clinic serving

children of that age range. It is desirable also to have tests of still narrower coverage, some for eight to twelve years and others for ten to fourteen years. Not until a test has been standardized will its exact coverage be known; but a test that has been made easier than was intended is more likely to be used than one that turns out to be more difficult than was intended. The standardization should include third grade for all sub-tests, and second grade for some of them.

The multiple-choice method is recommended for most tests, with four words from which to choose. The child of third-grade reading ability should not be given too heavy a task in reading. The significant word should be marked where it stands and when the child makes his choice. On no account should he be required to respond by writing a number, or to find another place for his response. Tests planned for mechanical scoring are not suitable for clinical use. Although it is not possible to arrange the items strictly according to difficulty, it is important to have a few easy items at the top of the page. The child of eight years should find something within his grasp, to keep him from becoming discouraged.

The "true-false" test is not recommended, because it permits so much play of chance as to necessitate penalizing the errors. The penalization is arbitrary and may introduce more errors than it corrects. Furthermore, it increases the danger of errors in scoring.

Weighting of items is also conducive to errors on the examiner's part. It is important that the examiner be able to score some sub-tests immediately, under cover, while the subject is working on the next task. Tests having uniform value for the items are much more easily scored than those having weighted values. Weighting is desirable for some tests, but it does not seem to add enough of value to the multiple-choice test to justify the effort.

A battery should contain several "information" tests, made up of questions in geography, history, literature, and science. One sub-test relating to current events might be useful for a few years, but there is danger that it might lose its value before being standardized. If used at all, items of current interest should be used in a test made up wholly of such items rather than with items that may be expected to retain their significance. In general, the information

test may contain items from miscellaneous branches of elementary school instruction. What the test measures, primarily, is the child's memory for instruction received at school and at home. How his memory for one topic compares with that for another school topic is a problem for the school rather than the clinic.

Tests that call for certain types of reasoning are more significant when treated as units and are more economical of time than tests that require shifting of the mental set. There should be several forms of "analogies," and numerous sub-tests calling for recognition of essential similarities and differences.

The "directions" test is one of the most useful written tasks, especially significant for showing the child's comprehension of what he reads. This test can be adapted to the lowest level of reading ability and can be used for children of seven years. At the upper age levels it can be made difficult either by timing or by introducing something unexpected. Unlike tests of the multiple-choice type, the directions test admits of partial credits. A convenient form of evaluation is to allow two points for a perfect response to any item and one point for a response that is at least half correct. A page having one item for each line is much more easily scored than a page in which the items are distributed through a paragraph of solid text. Being free from the influence of chance, the directions test may be very short.

The written sub-test, like the oral scale, should be passed upon by disinterested critics before being adopted for standardization. The preliminary forms may be mimeographed, always with items to spare. The ultimate criterion for any item is the response it elicits from the child examined in the clinic. The forms should be handed out freely to clinical workers who are willing to give the test a trial.

Two manipulative language tests may be included as written tests on the ground that they are for literate subjects. These are offered for the consideration of a school psychologist.

"Word building" is as simple a task as can be devised. The letters may be taken from the game of "Logomachy," or may be cut from papers and mounted on small cards. The number of letters should be not less than eight nor more than twelve, and there should

be no duplicates. The task is to make as many words as possible in five minutes, and the score is the number of words. All the examiner does is to list the words correctly assembled by the child within the time limit. This admits of group presentation, with a slight loss in accuracy. (Some children will spend more time writing their words than others.) The eight letters needed for the words cat, dog, and boy are recommended, plus two or three other letters if desired.

"Sentence building" is a similar task, with words instead of letters on small cards. Materials can be prepared by cutting words from advertising pages of magazines. (A toy printing press, if available, would be very useful.)

There should be a series of sentences, beginning with several of four or five words, such as "the cat has four kittens." The test would be discriminative at seven years, and yet could be carried up to a degree of difficulty appropriate for high-school students. A period may be included on a separate card, and a comma or two if needed. There should be no duplication of words in a given sentence. The words for each sentence can be kept in an envelope, identified by a number. A time limit is essential, to keep a child from working indefinitely on a task beyond him. The subject may be permitted to go as far as he can in the series, and a failure may be followed by presentation of a short sentence which has been held in reserve. The score might be the total number of words included in the sentences correctly assembled.

From the child's point of view, this test is a game having little in common with the "mixed sentence" test in the Stanford-Binet scale. The task of assembling words into a sentence is incomparably more interesting when the child is permitted to manipulate the materials. One disadvantage of the manipulative task is that it introduces the factor of chance. This makes it necessary to use a large enough number of sentences to offset the interference.

8. Criteria for Nonverbal Scales

Ever since the development of Army Beta (166) for use in World War I, there has been a recognized demand for a group-pre-

sented scale for illiterate subjects. The need has not yet been met, and it must be acknowledged that the essential requirements are not easily met. The failure of the various nonverbal scales to serve their purpose will be examined at greater length in a later chapter (pp. 134-137). It is sufficient here to state some of the criteria that should be observed in the construction of a scale for illiterates.

The test should be untimed. Group tests are necessarily standardized upon literate school children, because no other subjects are usually available for group standardization of a test. Illiterate persons should not be expected to compete with school children in the ease and speed with which they can handle a pencil. Let the test be standardized without time limit, so that the illiterate subject may be permitted to take as much time as he needs.

All pictures should be large, clearly drawn, and widely spaced. Eyestrain should be reduced to the minimum. Illiterates are not accustomed to using their eyes for close study of a page of tiny pictures. The task presents quite enough difficulty even when every picture is clearly recognizable.

There should be nothing obscure or confusing. An unfinished picture to be completed should be complete in every detail except the one that is to be completed by the subject. There should be no reading matter as a part of the picture, nor anything that would be puzzling to an illiterate.

Each page should be standardized as a unit. It is much easier to conform to these requirements for a single page than for an eight-page booklet. If the pages carry separate norms, an examiner may use those sub-tests which are suitable and omit the unsuitable ones.

The demonstration pages or lines should contain abundant material for each sub-test. The task should be the same for the whole page, so that the subject need not be interrupted in the middle of a page.

Tests thus developed and standardized would be of inestimable value in the examination of nonreading children.

9. Criteria for Standardization

The first thing to consider, when formulating a test with reference to standardization, is its essential standardizability.

Psychometric tests for children are under consideration here, not projective techniques. A higher standard of objectivity is required for a test to be used for numerical evaluation of a child's mental capacity than is either possible or desirable for a method of study developed by each examiner from his own experience. The writer uses many unstandardized tests for free interpretation, and encourages other students to use them so; but considers it highly questionable to use for numerical evaluation a test—however well standardized—that is not really suitable for standardization. The fact that the results have been collected in quantity and compiled in the form of statistical tables does not make a test standardizable in any significant sense.

Among the various nonstandardizable tests that have been standardized, the free-association test (69) occupies a very high place with reference to its nonstandardizability. Whatever value it may possess as a psychiatric instrument is due to its possibilities as a projective technique, not to its norms. The writer has not used this test more than twenty times since its publication in 1910, and does not recall having used the norms even once.

In a psychometric test depending upon language, every question or item should admit of a response that the examiner can recognize unfailingly as correct. There may be more than one correct response, but there should not be more degrees of acceptability than can be provided for by a not-too-complicated splitting of credits. Each question should be sufficiently specific so that an experienced examiner can pass upon the response with reasonable confidence. There should not be so much room for disagreement between examiner and author as to keep the examiner in a state of conflict over the fairness of the scoring.

Uniformity of presentation and evaluation is presupposed in a test yielding numerical results; and therefore the test should admit of uniformity in the highest degree possible. In this respect the written test possesses a great advantage over almost any orally-presented test. A multiple-choice test can be scored with complete objectivity; and a directions test can be scored almost objectively, although it permits greater latitude of response. The arithmetical test can be scored objectively, whether presented as an oral or a written test. Among orally-presented scales, the opposites test ranks high with reference to objectivity of evaluation. The original Binet-Simon scale and all the modified forms that stem from it

contain numerous items involving no subjective evaluation and perhaps permitting only one correct response; but they contain also items that elicit doubtful responses, thus vitiating the objectivity of the whole scale. It is recognized that the most revealing responses cannot be scored by a plus or minus; but the individual responses, however significant, do not permit any such certainty of scoring as do those responses that are undeniably either correct or incorrect. In the writer's opinion, the safest course is to use the individual responses frankly for subjective evaluation; and to base the score upon those responses that can be evaluated objectively beyond reasonable doubt.

The children employed as subjects for standardization of a test should be as representative of the population as can be found. A small city is preferable to a large one, because the public schools of small cities include a heavier proportion of the child population. The schools selected should be well balanced as between industrial and residential neighborhoods. The ideal school is one in which the grades are divided into three sections, according to achievement; and it is advantageous to draw as much data as possible from the middle-section children.

For an individually-presented test it may be well to select from each grade the children who are of the correct age for the grade, if it can be ascertained that all promotions are honestly earned. There is nothing to be gained by it in those schools in which it is customary to promote the children by age rather than by achievement. In any grade it is desirable to exclude children who are almost halfway between birthdays.

Of course there must be no arbitrary selection either of children for an individual test or of groups for a group test. All selection of subjects for standardization should be according to rule. A plan for selection should be formulated in advance, according to local conditions.

The standardization of a test should be carried one year in each direction beyond its true discriminative capacity. Group standardization is wasteful at the ends of the scale. The highest grade will include the retarded children of a given age, but not the bright ones who have passed on into the next grade. The lowest grade, on the

other hand, will include the brightest children of a given age, but not the retarded ones who have been left behind. Thus the ages at the ends of the scale are thrown out of balance. The only way to cover these ages properly is to obtain additional records beyond the grades for which the test may be expected to be useful.

A test intended to be used for individual study should be standardized with something of the latitude with which it will be used in the clinic. Formal presentation is out of place in an individual interview. Whatever rules are offered for presenting a test, the clinical examiner is practically forced to take some liberties with them in order to obtain responses that represent an approach to the child's true capacity. Similar liberties should be taken by the students who collect the data for standardization, so that the norms may be the more applicable to the child examined in the clinic. Inasmuch as the test will be used by many clinical examiners of diverse training, it is desirable that there be a wide representation of students and universities in collecting the data for norms. Strict uniformity of presentation does not necessarily render the method objective, but may serve rather to introduce inaccuracies in its application.

For tests presented individually the informality of presentation will perhaps take care of itself. It is for the group-standardized test that this should be kept in mind. No rules should be laid down for group presentation that cannot reasonably be followed in an individual interview, such as instructions for "Pencils up." The task should be explained in natural diction, not according to a formula that has been memorized verbatim. Sample items may be presented from the blackboard before the test forms have been given out, inasmuch as it is much easier to hold the attention of the children while their desks are clear. Following the explanations, the test forms should be laid on the desks with the blank side up, and the children should write their names on the back side of the sheet or set of sheets. All timing should be for the page, not for the individual line. For timed work, the two-minute interval is recommended as being long enough to be measured with fair accuracy but not so long as to be fatiguing. After the timed work for each page has been completed, the children may be allowed to finish at their

leisure with red pencils. Corrections may be permitted with the red pencil, but there should be no erasures. Each set of test forms handed in should be inspected to be sure that no page has been overlooked on the second round.

Each page yields two scores: the timed score including all correct markings made with lead pencil and the untimed score including all correct or corrected responses on the page. Two full sets of norms can be derived from the records, but the timed performance may be regarded as a by-product. The norms based upon untimed achievement represent the real purpose of the standardization, because comparatively few of the children examined in clinics can be rated fairly by timed tasks. It is, however, of considerable interest to be able to compare a child's timed and untimed achievement. Inasmuch as the timed scores can be obtained with a very little additional work, it is well worth while to present the tests first with the time limit. The timing is of special significance for the older children included in the standardization and is comparatively unimportant for children below fourth grade.

The children should be grouped according to age nearest birthday. The age norms should be for the year rather than for the half-year. Records from children whose ages are almost exactly on the half-year may be discarded.

For most tests it requires many thousands of cases to produce a smooth age curve. In general, it may be expected that the curves will require a little artificial smoothing, but no one should publish norms thus derived without publishing also the true figures. Anyone who accepts and uses an artificially corrected age curve has a right to know the exact nature and degree of the alteration.

10. Validation of a Battery

In the standardization of tests as in their use, there is a vast amount of "straining at gnats and swallowing camels." The more heavily we rely upon statistical "validation" of tests, the easier it becomes to overlook what makes a test truly significant and helpful in the study of the child who is referred to the clinic. If the child's cooperation cannot be obtained in the examination, any claim for

the statistical validity of the test is hollow—the cooperation, not of the hundreds of children from whose records the norms were derived, but of the particular child whose record is evaluated by these norms. One child can *in*validate an examination.

Very rarely can an observant clinical examiner state with confidence that an examination represents the child's full cooperation in every detail. More often it is easy to name certain items in which his cooperation falls short of being adequate. There is usually no witness who can support or challenge the examiner's impression that the examination is satisfactory; and the examiner's fallability in estimating the degree of the subject's cooperation has already been noted—and acknowledged.

No form of examination will insure success in every case, because the responsibility for its success rests in part upon the subject. The reaction of the subject is unpredictable, and it is safe to assume that there will always be cases in which the only honest report to be offered is the doubtful one. In the large majority of cases, however, the child responds well to some part of the examination. The "flexible" scale would not solve all problems that arise, but it would go far in helping the examiner to gain control of the situation. If there were enough sub-tests to draw from, a tactful examiner would usually be able—after a little trial-and-error work with nonstandard tasks—to find something that would hold the child's interest. Any tasks in which he refuses cooperation may be thrown out, and the score can be based upon those tasks that are believed to represent satisfactory cooperation.

Of course, the tasks thus selected by the subject would not necessarily constitute a well-balanced battery. If the subject were to be excused from all tasks not to his liking and rated wholly on those having a strong appeal for him, the battery might offer nothing more significant than a dozen slightly different ways of measuring the same aptitude. It is essential to the adequacy of the examination that various aptitudes be measured; and the battery can be accepted as well-balanced and adequate only when the ratings by the various sub-tests show a considerable spread.

All available sub-tests should be classified according to their intercorrelations; and those tests that seem to measure the same

aptitude should be grouped or paired, so that the examiner would know where to find the best substitute for any particular sub-test that had been omitted or discarded. The value for this purpose of an exhaustive study of intercorrelations among the sub-tests can hardly be overstated. However, inasmuch as the battery is intended for study of individual achievement, this writer would like to see it put to a more severe test than can be furnished by statistical criteria: it seems only fair to demand that the battery yield passably uniform results from a small group of normal children known to be well-matched in their school achievement.

No particular sub-test could be expected to bring identical scores even from a pair of identical twins; but it should be possible to make up a series of tests which, taken as a whole, would show some approach to uniformity when applied to a group of children selected as being approximately equal in their school achievement. If not, the propriety of using tests for estimate of any individual child's mental capacity may well be questioned.

Only in a carefully selected school could this test be made with any reasonable chance of success. It should be a large school in a home-owning suburban community, and the assistance of the school psychologist should be sought at least a year in advance. The school should be one in which it is customary to divide the children into sections in accordance with test findings; and the middle section—preferably the middle of five sections—should include no children either retarded or advanced. The test should be made in the spring term, and the test group should be made up of children who had been under the same teacher all the year. There should be no arbitrary selection of individual children; but it would be permissible to restrict the group to those of approximately the same age, and to exclude both those who had lost time because of illness and those who had entered in the middle of the year. The group would of necessity be a small one. The number of sub-tests used should be at least twenty—all language tests and all so standardized as to yield mental-age ratings. They should include oral tests for individual presentation, and written tests group-administered with no time limit. The tests should be as varied as possible with reference to the type of verbal aptitudes measured by them, but reasonably uniform

as to the time required. There should be enough examiners to permit all oral examinations to be made during one session, so as to allow no opportunity for coaching.

It is useless to try this experiment without having full control of the conditions. First, there should be exhaustive effort to produce sub-tests in which normal children of a given age may reasonably be expected to achieve a passing score; and then no effort should be spared to make the test to which the battery is to be subjected as fair a test as possible.

In the later 1920's the writer made a very crude attempt to employ this principle as a check on a seven-unit battery that had been tentatively standardized(73). The battery did not pass the test, as might have been foreseen. The number of sub-tests was too small to allow any margin; and the subjects included all the twelve-year-old children from two schoolrooms. All the sub-tests used have since been either discarded or revised, but the failure of the experiment was not due wholly to the inherent weakness of the tests.

The results of this experiment would be significant, whether positive or negative. One could classify and reclassify the mental age ratings of each child by each test in as many ways as desired, without once losing sight of the individual child. If it should be found possible to select a dozen tests that would give each child a mid-score rating close to his actual age, these tests would furnish material for a standard battery. The remaining sub-tests, possibly excluding some showing anomalous results, would be available as alternates, each to be paired with the standard sub-test which it most closely matched.

On the other hand, if no series of sub-tests could be found that would yield approximately equal median ratings for all the children, the negative findings would be extremely important as showing the weakness of any claim that statistically established norms can be fairly applied to individual achievement.

11. Possibilities of the Formboard

Of all tests, language and performance, the formboard comes nearest to being universally applicable. Provided it be of difficulty

appropriate to the subject's level, it has an appeal that is all but universal.

More formboards are needed, especially at preschool levels. The development of new formboards is a task for a student who has access to a motor-driven scroll saw and who possesses considerable manual skill. Mechanical aptitude of a high order is not necessary nor even desirable. A person having a too-ready grasp of mechanical problems is likely to invent tasks too difficult and too puzzling for serviceability in the clinic. Very simple tasks are most needed; and patience on the part of the designer is needed to make them simple enough. It is of special importance to avoid anything tricky or confusing.

The board for preschool children should be about ten inches wide—never more than twelve inches. The blocks should be of a size that can be handled comfortably by a child's hand, preferably a little larger than the blocks of the Witmer board (167, 170). The number of blocks for a given board should not be greater than can be made distinctive in form, without the introduction of any complex feature. The blocks should be at least one-eighth inch thicker than the depth of the recess; but one may make one cutting serve for both board and blocks by inserting a false bottom in the recess. The blocks should be of solid wood, about three-eighths inch thick; and the board, if cut separately, may be of one-fourth inch plywood.

The task for the young child should be self-corrective. This can be accomplished by having the blocks approximately equal in size but differing in form. It should be impossible to insert any block in any recess except its own, but each recess should be large enough to take its own block easily. In order to get the figures sized so as to exclude incorrect placements, one may cut cardboard patterns of the blocks and match each one with each of the others.

The triangle is a poor figure for the preschool test. For some reason not easily explained, this figure presents more difficulty than the semicircle. The two figures are alike in that either block may have to be turned around to make its base match the base of the recess; but the child usually sees what to do with the semicircle, whereas frequently he tries to insert the triangle with the base at the top.

The standardization of a formboard for young children should be such as to allow credit for an incomplete performance. The child who places all the blocks may be given a score based upon the speed of his performance; and the score of an incomplete performance may be the number of blocks placed. The range of discriminative capacity is widened by recognition of the partial performance, and the test is thus made to cover the level at which it is most useful. There are many small children who will place a few blocks and then become discouraged.

Six simple formboards having from four to ten figures would meet a very urgent need of the clinic, with or without norms. They should differ in size and design, as well as in difficulty, and should be presented in such a way as to minimize the effects of practice from one board to another. Practice effects may be studied by presenting the same board twice. Whatever the outcome of a repeated performance, the second trial is more significant than the first. The boards will be more useful if standardized singly than standardized as a series. It cannot be known in advance how many tasks a preschool child will elect to perform. In presentation it may be desirable to start with the most difficult board, and to offer an easier one in the event of failure.

For older children, however, a series of increasing difficulty is more serviceable than the single task. It is desirable to have such a series discriminative for children from six to twelve years, but the need is not quite so urgent for school children as for preschool ones.

Development of upper-level formboards is less rewarding, because they are too costly to be widely used. Increasing difficulty of solution increases also the difficulty of manufacture. Several formboard series are available that are graded up to the adult level. They are not perfect, but it is their cost rather than their defects that limits their usefulness.

12. Development of Color-Cube Scales

The Embossing Company's Color Cubes have been used by many students for diverse purposes; but the possibilities of this compact and convenient piece of apparatus have by no means been exhausted.

There is still a place for two additional scales: a lower-level scale employing only the solid-color sides of the blocks, and an upper-level scale requiring less time than any scale yet standardized.

In 1923, after having found that the Kohs Block Design Test (92, 93) required as much time as could be allowed for the entire examination, the writer undertook to shorten the series for personal use in the clinics of Worcester State Hospital (78). Inasmuch as no use could be made of the Kohs norms for a shortened series of tasks, there was nothing further to be lost by making other modifications that seemed desirable. Accordingly, the designs were built on the same scale as the cubes; and the score was based wholly upon the time records, on the ground that the move-count is too inaccurate to be used in evaluation. Some new designs employing solid-color blocks were added to the series, but there was no attempt to avoid using the Kohs designs. Five of them were adopted exactly as used by Kohs, except that they were enlarged; and two others were altered only in color.

When devising a modified form of a test published by another, one should make the new form as different as possible from the older form, in order to permit both forms to be used for the same subject with minimal interference. It is well to keep this in mind even when developing a test for personal use, because its future cannot be foreseen.

The Kent edition of the Kohs test was not originally intended for publication or standardization, but it was given out privately to others who found the Kohs test too time-consuming. After the series had been standardized for personal use, it was published for the convenience of those who were already using it. The encroachment on the Kohs test has been a matter of regret to the writer for many years. It was entirely unnecessary, because the cubes permit many more designs than have even yet been used. There are still enough unused designs for the two new scales described in this chapter.

The proposed lower scale will be referred to as the "square" series, because each cube used in any design is to be a solid square of a given color. The upper scale may be called the "triangle" series, on the ground that each design is to consist largely of red-white or blue-yellow triangles. The solid-color sides of the cubes should

occupy a minor place in the triangle series, without being wholly excluded.

The degree of difficulty is determined rather more by the symmetry of the design than by the number of blocks, the symmetrical design being more easily comprehended than the irregular one. A symmetrical design should be used at the lower end of each scale, and the tasks at the upper end should be asymmetrical.

The second design used in the Kent series turned out to be much more difficult than was intended. It is design D, an irregular figure consisting of five blue squares and four yellow squares. It is now believed that this design involves an optical illusion, which is not apparent to the practiced subject. Nearly all the children make the same error: they start correctly with a yellow square in the upper corner, and then build three blue squares onto it instead of two. They seem to overestimate the size in the horizontal direction, and occasionally in both directions. Many children who fail helplessly with this nine-block design are able to perform with ease the sixteen-block task that follows it. This difficulty, due to irregularity, should have been reserved for the final task.

Each series should be planned with reference to the lower end of the scale, to bring its discriminative capacity down as low as possible. It is much easier to devise color-cube tasks sufficiently difficult for high-school levels than to make tasks easy enough for preschool children and for retarded children. It is at the lower mental levels that tests are most needed; and therefore the scale of "square" designs is more important than the "triangle" series.

The full-size design is preferred to the half-inch scale used by Kohs (92) and Wechsler (155), because the larger design is easier to comprehend at the lower levels. The difference in size between design and blocks is unimportant for the subject who has a thorough grasp of the problem, but it is confusing to the child who makes hard work of the task at best.

The test cannot be used economically in a continuous series covering all ages. It would be a waste of time to standardize the "square" series for upper-level subjects, and impossible to obtain norms for the "triangle" series at the lower levels. The two scales should over-

lap in coverage, but not in content. In most cases only one scale would be presented, but there would be no objection to using both scales and obtaining two scores within the range of the overlap. It is probable that both scales can be made discriminative for the ages seven to nine years.

The designs for the early tasks of the "square" series should be built up of blocks, not painted on a card. These tasks need not be timed. A correct performance, unaided, should receive at least one point credit regardless of the time required. The first performance from a painted design may also be credited without timing, after which the performance may be scored by speed.

At the upper end of each scale the subject who passes the time limit without completing the task may receive credit for the number of blocks correctly placed.

13. General Criteria for Performance Tests

The problem for preschool levels is to have a sufficient number of tasks, sufficiently varied in kind. For a task of whatever type, the range of difficulty should cover the mental levels from two to six years. Tasks that can be scored without timing are especially at a premium. Detached tasks are more serviceable than the graded series, because of the uncertainty of a young child's cooperation. Provision should be made whenever possible for scoring an incomplete performance. The method of scoring should be adapted to the age for which the test is to be standardized, although it is likely to be used more frequently for older children of retarded development.

The toy shop is the place in which to look for test materials. The color cubes, first sold as a toy, have scarcely a rival for general serviceability as a clinical test. By visiting toy shops during the Christmas season, someone may discover another set of constructive materials holding possibilities for definite tasks. The toy which permits different forms of construction with different degrees of difficulty is something to be acquired and "played with" in every possible way.

For children of school age the graded series is to be preferred to the single task. The influence of chance is a disturbing factor

in manipulative tasks, as was clearly apparent more than thirty years ago when the Pintner-Paterson scale (121) came into use. The effects of chance may be minimized in the score by having a series of tasks that can be graded up in difficulty by quantitative steps. It is true that the effects of practice may produce an unexpectedly high score, but this need not be considered as interfering with the natural course of performance. The subject is honestly entitled to any help he receives from practice; and his ability to take advantage of practice is one of the most important findings to be observed. He may forfeit the score by exceeding the time limit, but the use he makes of his experience in the solution of a similar but more difficult problem is not the less significant. It is the graded series, rather than a collection of detached tasks, that offers the best opportunity for observing practice effects.

The defect most commonly found in performance tests is some feature that is unduly puzzling. The "Passalong" test (1) is attractive to a subject who has a taste for mechanical puzzles, but annoying to one who cannot grasp the principle upon which it is based. The Vigotsky test (55) may keep the subject working many minutes on false leads, and the chances are against a child's finding the correct lead at all.[1]

In experimental studies the value of such tests is not to be questioned. Industries and vocational schools have need for tests that measure distinctive and highly specialized aptitudes. Quick comprehension of obscure relationships is something well worthy of being measured; and the ability to shift rapidly from a given plan of work to something entirely different is obviously valuable in certain fields of endeavor. Tricky mechanical puzzles may be serviceable for some purposes, along with the "number series" test and other problems that strongly suggest false leads. The subject who can resist the suggestion possesses an aptitude of undoubted value, however far removed from what we call "general intelligence."

[1] A new test which was being used experimentally with a few highly intelligent subjects elicited the following comments from a woman who was conscientiously cooperative: "If I only knew what I am striving for." "There doesn't seem to be any reason for anything." "I don't know whether it's for a balanced design, or what." "Nothing is related to what goes before, there doesn't seem to be any key."

The development of tests for measuring this aptitude should be encouraged, but it does not follow that the tests should be standardized for children. The tests most needed in the clinic are those that can properly be used for the majority of children rather than those suitable only for the exceptional child. In most cases the test that is conducive to a sense of confusion and frustration is worse than useless. The child's reaction of helpless bewilderment may be carried over to any tasks that follow and may tend to invalidate the examination as a whole. The most serviceable performance tasks are those involving the basic aptitudes that any normal child may reasonably be expected to possess, in degree appropriate to his age.

It is from the child's achievement that something is learned rather than from his failures. It would be easy to make the examination so difficult as to insure total failure, but the findings would not be useful. The real purpose of the task is to find out what the child can do at his best; and the strongest performance test for clinical use is the one offering the largest number of children the opportunity for some measure of success.

CHAPTER IX

SERVICEABLE SCALES AND SUB-TESTS

Introductory Note

The recommendations here offered are based upon personal experience, influenced as it is by lack of contact with what other students in the field are doing and by the inevitable tendency to use a well-known test in preference to a strange one. The collection of tests described in this chapter is not offered as being exhaustive, nor is there any claim that it includes the best of the tests now available.

The Buros "Third Mental Measurements Yearbook" (23) is a gold mine of information about the tests developed since 1940, and should be in the hands of every psychologist who wishes to keep his equipment up to date. The new test material offered for children, however, is somewhat disappointing as compared with that available for adults and adolescents.

The tests developed by the writer are referred to repeatedly as having been devised specifically for examination of children in the clinic, but it must be kept in mind that they have not been properly standardized; also, it must be acknowledged that most of them are not now considered worthy of further standardization in their present form.

The large majority of tests are imperfect to begin with and are inevitably in a state of depreciation. The adoption of any test or system should be tentative, not final. It is only by being always on the alert for new developments that any examiner can make the most of such resources as are available at any given time.

1. The Detroit Tests

It has already been stated that there is no complete scale which can be recommended in its entirety for an individualized examina-

tion. The scale that has most in its favor is the battery known as **Detroit Tests of Learning Aptitude**, by Baker and Leland (10). The essential materials for this scale are compact and moderately priced, consisting only of a small handbook for examiners and a spiral-bound book of pictorial material for the subject. This outfit should be included in the equipment of every clinic, even if the scale is not to be used for the whole examination. Every clinical examiner should master the technique for presenting a few of these tests, perhaps giving preference to those that permit the results to be recorded on plain paper rather than in the record booklet.

The scale includes nineteen sub-tests, each standardized as an independent unit. The set-up of the scale affords all the flexibility that could be desired, inasmuch as the examiner has utmost freedom in his choice of sub-tests. It is the sub-tests, not the scale, which impose rules not willingly obeyed by the examiner who wishes to show all possible consideration for the feelings of the subject. These criticisms serve well to illustrate how easy it is to spoil a test possessing some very strong features.

Most of the sub-tests have a wide range of coverage, some of them extending from three years to the adult level. Several of these are to be presented in full for a subject of whatever age; and for others it is required that there be a specified number of failures at the "ceiling." It follows inevitably that the time is used wastefully and that the items having no discriminative value are unduly fatiguing to both subject and examiner. Also, the responses elicited by many of the questions are so difficult of evaluation as to require constant reference during the presentation to an elaborate and somewhat arbitrary scoring key, thus making it very difficult to find the subject's ceiling. Another defect, one that might easily have been avoided, is that some of the pictures in the test "Pictorial Absurdities" contain a little reading matter essential to the meaning of the pictures, thus rendering that sub-test useless for a child of reading age who cannot read.

In this scale, as in any flexible system, the sub-tests must be selected with care. Only those who wish to "have their cake and eat it" will object on this ground to the flexibility of the system. With all its imperfections, the scale contains serviceable material. In a

clinic in which the routine use of the Stanford-Binet is required, one or two of the Detroit sub-tests might be of great value as supplementary measures. The test called "Orientation" is especially useful at the lower end of the scale, and might well be used regularly for preschool children whose ceiling would be reached within the first twenty-four items. The rest of the series is unduly time consuming, because of the frequency with which the subject is required to leave his seat. There are occasions when it is desirable to have tasks which permit a child to move about the room; but it is the examiner, not the author of the test, who knows when such a task can be used to advantage. The test called "Oral Commissions," discriminative up to eight years, can be used most appropriately for the child who is too restless to keep his seat; but it is not recommended that it be presented while a child is quiet and attentive.

Other tasks of this scale, selected according to the examiner's taste, will certainly be found useful on many occasions.

Two group-standardized tests which should be kept in stock are **Detroit Beginning First Grade Intelligence Test,** by Engel and Baker (38), and **Detroit Primary Intelligence Test,** by Baker (9). The pictures in these tests are clearly drawn. It is permissible to use either of them for presentation to a small group.

The **First Grade** test covers the second-grade mental level as well as first grade. Norms are offered for the ages four to nine years, but the test cannot be considered serviceable for children under five years or over seven years. The booklet contains ten untimed tasks, all included under a single score. Although it would be convenient occasionally to use the test in part when there is not sufficient time to present it in full, the objection to the composite score is in this instance theoretical rather than practical. The test is especially useful for a six-year-old child who has adequate understanding of language but who does not respond well to questions.

Detroit Primary is for grades two to four and is useful up to the age of ten years. This booklet contains seven sub-tests, having independent norms. Unlike the **First Grade** test, the **Primary** test depends for its serviceability upon having separate norms for the different sub-tests. It is the first three pages, which differ only in the degree of difficulty from three sub-tests in the **First Grade** test,

that are useful in a clinical examination. One of the four sub-tests not recommended is a page of questions to be answered by underlining "yes" or "no," with no provision in the scoring for counteracting the influence of chance; and another is a "number-series" page, a test that does not appear to have any value for children below high-school levels. The last page in the booklet has been found useful for a purpose quite different from that for which it was intended.

This test depends essentially upon a child's recognition of capital letters, as compared with a series of larger capitals at the top of the page. The task has something in common with the "symbol-digit" test and involves considerable eyestrain. It has been found serviceable for the nonreader, because it serves to indicate whether his inability to read is or is not due to his failure to recognize the letters. Occasionally a subject who cannot read one word performs this task more quickly than does one who reads normally. This may be explained on the ground that the literate subject's attention is distracted from the task of matching letters by an irresistible tendency to look for a word in each group of letters; whereas the subject who cannot read is able to give his full attention to the letters as such, and does not waste any time noting that the letters of a given group almost spell a word. Some nonreaders, however, cannot recognize even the letters and cannot see the similarity between two A's of different size. The clinical examiner may not be prepared to make an analysis of the type of a child's reading difficulty, but it is of some significance to ascertain that he can or cannot match the letters. For any nonreader, the results are worth the three minutes occupied by the test.

The **First Grade** test is recommended for a child whose mental level is below eight years, and the first three pages of the **Primary** test may be used for a child of eight to ten years. The last page of the **Primary** test may be cut off and saved for a nonreader.

Detroit Word Recognition Test, by Oglesby (113), is included here as bearing the name "Detroit," but it is not otherwise related to the tests described above.

This is not a psychometric test, but it is valuable as a quick means of ascertaining how a child's reading ability compares with

his mental level and his school placement. Having a time limit of four minutes, it does not require more time than can be spared from the routine examination. The results may indicate that a child is making normal progress in learning to read, or may serve to show that he should be referred for special examination in reading. In the case of a mentally retarded child, they may indicate that he is not likely to acquire reading knowledge and that his time can be spent to better advantage on manual training.

The pictures in this test are beautifully drawn and are used very lavishly. The task is to draw a line from a word or group of words to the picture named or described. The first page of the four-page leaflet is for demonstration. On the second page there is a picture illustrating each word, and for the last two pages the reading matter for a given picture is increased up to seven or eight words. Significant words are used several times on a page, with remarkable ingenuity, furnishing many pitfalls for the child who reads one word and guesses at the meaning of the phrase; and thus there is little danger of a correct response by chance. The task is fascinating, whether for the child who can read only a few words or for one who can earn a perfect score.

There are four forms of this test, and at least two forms should be kept at hand. The number of the form used should be noted in the record, so that a different form may be used for a repeat examination. The four forms are essentially equivalent and are evaluated under the same grade-norms. Inasmuch as the norms are not in convenient form for ready reference, the following table is offered. (Anyone using it should check it by the data published in the manual.)

NORMS FOR DETROIT WORD RECOGNITION

Grade	Score
Low first	7—12
High first	13—19
Low second	20—25
High second	26—30
Low third	31—35
High third	36—40

It would be a waste of time to use this test for a child of higher than third-grade reading ability, and there is some danger that he might resent it as beneath his dignity. When in doubt about the appropriateness of this test, the examiner may present the easiest of nonpictorial written tests with the request that the child read the first line aloud. If he requires any help in reading this line, it is obvious that the test is too difficult for him. In this event, **Detroit Word Recognition** should be substituted as quickly as possible, to see if he can read the words on the sample page. If he cannot read this page unaided, he should be allowed to think that the sample page is the entire test and should be given enough help to enable him to point to the correct pictures. The leaflet should not be opened until the examiner is sure that the child can read well enough to earn a score on the test, and the pencil should not be offered unless it is to be used. For any child who can read the words on the sample page without assistance, it is safe to assume that the test will reveal enough to make it a good investment of time.

The findings by this test are of special significance with reference to the child's placement in school. Very frequently the results show that he has been promoted beyond his reading level. Occasionally one can learn from the mother that she had insisted upon his being promoted, against the advice of the teachers.

The psychometric findings in such cases are of great importance. The child who is mentally retarded may quite possibly make better social adjustment in school if permitted to sit in the room with children of his own age than if kept in the grade with smaller children. If he is destined to grow up illiterate in any event, it will do him no serious harm to let his promotions be determined by his age and size rather than by his achievements. Perhaps the mother may be advised to read aloud to him as much as possible, and to teach him to learn by listening. When a child lacks the capacity to acquire reading knowledge, it is more important to teach him something that he can learn than to keep him struggling for years with a reading book he cannot master. On the other hand, the child of near-normal mental level should be taught to read if it can possibly be accomplished. This child's mother should be urged to seek expert

advice on remedial reading and to use all her resources for teaching him to read. It should be pointed out to her, with tact and patience, that his time in school will be largely wasted if he cannot keep up with his class in reading, and that he can well afford to spend an extra year on this one topic.

The number of nonreaders referred to the clinic is large enough to indicate the need of special equipment for analysis of reading difficulties. **Detroit Word Recognition** is a very satisfactory test as a preliminary measure and is especially valuable because of its simplicity. One need not be trained in remedial reading to make use of it.

2. The Kuhlmann Tests

In the writer's experience, the **Kuhlmann Tests of Mental Development** (100) constitute the best scale for preschool children. The technique of presentation and scoring offers some difficulties, but it would be a valuable piece of equipment for any examiner who works with very young children sufficiently often to keep in practice.

Above the preschool level the scale cannot be recommended. Most of the school-age tasks are scored largely by speed, which impairs the value of the scale for clinical use; and the numerous written tests render it wholly inapplicable to nonreaders. There are eighty-nine tests, arranged in order of difficulty in a continuous series from infancy up to adult levels. Nineteen alternates are provided, but in all other respects the procedure is rigid. Each examination extends from the "base" to the "ceiling."

This scale has been described by five reviewers in the Buros 1940 Mental Measurements Yearbook (22), and it is significant that there is so much to be said about it. There is very little overlap in these reviews, written from different viewpoints.

The Kuhlmann-Anderson Intelligence Tests (97), a series of thirty-nine sub-tests having independent norms, include many of the tests used in the Kuhlmann scale. The coverage of the series is from six years to "maturity," and the tests have been group-standardized through the twelve school grades. The manual of instructions is very convenient to use, having a copy of the test itself

on the page opposite the page of directions for presenting that test. The forms are published in booklets for use in schools, and also in separate sheets for use in individual examinations.

The picture tests at the lower end of the series are timed line-by-line, with fairly liberal time allowance. Many of the pictures are too small to be easily recognized. These sub-tests for young children do not compare favorably with the **Detroit First Grade Test,** but they possess the great advantage of being so standardized as to permit their being used singly. For a child of six or seven years it is well to give preference to the Detroit test whenever time can be spared for presenting the eight-page booklet, and to use two or three of the Kuhlmann-Anderson picture tests when time is limited.

The sub-tests at higher levels are timed by the page, having time limits of from one to three minutes. For children who can read and who do not react unfavorably to being timed, the battery offers much that is useful. The "number-series" units, of which there are several at different levels of difficulty, are not recommended; nor are several other sub-tests, which seem to measure nothing more significant than ingenuity in purposeless manipulation of numbers and letters. These tests, however undesirable, are entirely harmless in a system that leaves the examiner free to choose what tasks are to be used. The series contains some serviceable sub-tests, and it is a battery with which every examiner should be familiar.

3. The "Emergency" Tests, Oral and Written

After the "shock absorber" tasks have served to gain a child's attention, it is the writer's custom to open the examination with one or another of the **"Emergency"** scales (88). In the large majority of cases one can find the appropriate scale without using any unnecessary questions and can in the minimum of time make a rough estimate of the subject's mental level. The findings of so short a scale should never be accepted as anything more than a preliminary estimate, but they may be helpful to the examiner who is uncertain at what level the formal examination should be started. The rating may be used also as one unit of a series.

The **"Kent battery of short written tests"** (89) is the fourth edition of a series of tests first formulated and tentatively standardized in 1925 (73, 83).

The series includes four multiple-choice tests: **Information; Similarity** (a vocabulary test); **Essential Property**; and **Essential Difference.** These tests are untimed and can be used for a child who reads somewhat laboriously. They are discriminative from eight to fourteen years, and are most useful for children from ten to twelve years. **Arithmetical Reasoning** may be scored with a two-minute time limit, and also with unlimited time allowance. All these tests are too short and too limited in scope to be used singly. They are scored singly, but intended to be used collectively.

Easy Directions is the easiest written test that the writer can name. It has been in use for fifteen years and has been found serviceable for ascertaining whether the child is a suitable subject for written tests. When in doubt concerning the subject's reading ability, the examiner may present this page with the request that he read the first line aloud. (**Detroit Word Recognition** should be within reach, to be substituted immediately in case the child is unable to read that line.) The subject who reads the line without difficulty is given a pencil and instructed to do exactly as it tells him to do, for the whole page. There are only ten lines, and they are so well separated that the examiner can observe the entire performance from the opposite side of the table. One who is familiar with the page can read it upside down, and can see which lines present difficulty of comprehension. Frequently the child reads the whole page aloud, although this is not requested except for the first line. It is sometimes observed that a subject who reads with ease makes several errors which are evidently due to failure of comprehension.

This test was tentatively standardized in 1933, and carries "mental-age" norms for the age range seven to ten years, but the norms must be used with caution. The significance of a perfect score depends upon the child's achievement in other tests. For one whose rating by other tests is above ten years, it would be unfair to include this ten-year rating in the series. A score short of perfect may be accepted as it stands, regardless of the child's achievement in other

tests, because his failure to earn a perfect score in this test proves that it is discriminative for him. He may be held responsible for an error in the test, but not for the fact that the test is too easy to cover his mental level. When the other tests prove that his achievement as a whole is not over ten years, a ten-year rating by this test is as good as that rating by any other test. It is when a perfect score on this test would lower his median rating by the series that it should be excluded. The examiner should assume full responsibility for having presented a too-easy task and should not let it be counted against the child.

One other test is included in this battery—**Graduated Opposites.** This is used as a written test for any subject who responds better to written tests than to oral questions; but it is more economical of time when used for oral presentation. There are thirty items, which form two twenty-item scales according to the plan described in Chapter VIII. The middle section of ten words is invariably presented first. A single error in this section may be accepted as proof that the lower scale is discriminative for the subject. The upper scale is rarely used, but is available for a superior subject.

The Woodworth-Wells **Hard Directions** test (161) may be mentioned here, although it does not belong technically to the "Emergency Battery" (Obtainable from C. H. Stoelting Co., No. 37031-G). First published in 1911, it was standardized by the writer about twenty years later with the third edition of the written battery. Two sets of norms, with and without time limit, were published in 1934 (77). The test is very attractive to some children, but should not be used for a child who is inclined to be suspicious. A normal child may be delighted when instructed to write a wrong answer to a question, and may perhaps make his answer as far from right as the space permits; whereas the child examined in the clinic may think it is a trick and refuse to write any answer. The test covers the age range from ten to fourteen years and is serviceable on many occasions.

There is real need for a directions test intermediate in difficulty between this **Hard Directions** test and the **Easy Directions** test described above, but the writer has not been able either to find or to

devise one of the desired degree of difficulty to be used as an untimed test. If the directions are made easy enough to be followed by a child of eight years, they are too easy for a child over ten years. A time limit may make the test discriminative for older children, but not without introducing a new factor.

4. Vocabulary Tests

Most of the vocabulary tests available are for high-school and college levels. Thirteen such tests are reviewed in the Buros Third Mental Measurements Yearbook (23). Anyone interested in tests for adolescents of the higher mental levels should look up these reviews. Only those that appear to be applicable to children in the elementary grades are mentioned here.

The **Shipley Vocabulary Test** (134) was formulated for study of mental deterioration; but inasmuch as it has been standardized for children upwards of nine years, it may be used for children of the fourth grade. A multiple-choice test of forty items, it is about the right length for general clinical use. The time limit is ten minutes, and this is so liberal an allowance as to make it almost an untimed test. In some cases the time limit is an advantage as a means of keeping a very slow worker from spending too much time on the test.

For a child of less than ten-year mental level, the **Similarity** sub-test from the Kent battery may be used as the vocabulary test. This has only eighteen items and should rarely require more than three minutes.

The **Holley Sentence Vocabulary Scales** (60) are offered in several forms for different mental levels. Each form contains seventy items, making a rather long sub-test for children below the eighth grade.

The battery **National Achievement Tests** (109) offers a vocabulary test of fifty items for grades three to eight. The writer is not personally familiar with this instrument.

No vocabulary test of a few hundred words can be recommended for estimating a child's total vocabulary. The sampling is too small to serve that purpose, and the child's score is inevitably affected by

his confidence and his willingness to guess at a response. All that can reasonably be expected from a vocabulary test requiring from ten minutes to an hour is that it show how many words of a given list the subject recognizes and how his knowledge of these words compares with that of other children of his age.

At least two picture-vocabulary tests have been standardized for preschool children, requiring no response from the child except that he point to the picture illustrating the word spoken by the examiner. One of these is the **Van Alstyne Test** (150). There are forty-five cards, each divided into quarters and having a picture in each corner. In presenting each card the examiner names the word and requests the child to point to the object or the activity named. Unfortunately, the pictures are too small to be easily recognized, and are not sufficiently attractive to hold the child's attention. The test is discriminative up to five years, and is strongest at the four-year level. It is permissible to divide the scale by using only the cards bearing odd numbers. In the majority of cases, the examiner will do very well if the child can be induced to look at the pictures for half the series.

The other picture test is the Ammons **"Full Range Vocabulary Test"** (3), intended for subjects of any age who cannot speak. The writer has examined this test, but has not used it. Although it does not appear to be very promising for preschool children, it may be useful for nonreading children of higher age levels. Unfortunately, crime occupies a rather important place in the pictures.

5. Language Tests for Adolescent Levels

The mentally retarded adolescent, who is frequently extremely sensitive, falls into the category with younger children with reference to the selection of tests. The tasks that can be presented with the minimum of discomfort to the subject have already been described.

There are many situations in which the **Wechsler Scale** (155) is the best available one, although it falls far short of the criteria enumerated in the last chapter and although it bears evidence of hasty work in the formulation of its sub-tests and items (82). When

the IQ report is required, as is common in examinations conducted primarily for the court, the Wechsler-verbal IQ can be recommended as being essentially more sound than the IQ rating derived from the mental age, by whatever test. As between the Wechsler and the Stanford-Binet scale, the Wechsler-verbal possesses important advantages for almost any subject whose mental level is at least ten years. (As already stated in Chapter V, the writer believes that the Wechsler scale is weakened, not strengthened, by its performance scale.)

For teen-age subjects of normal mental development, there are many group-standardized school tests from which any examiner may take his choice, according to his taste. Since 1922 the writer has used **Terman Group Test of Mental Ability,** 1920 edition (145), and has failed to find any satisfactory substitute among tests of more recent development.

This test consists of ten one-page sub-tests, the task for each page being relatively homogeneous as to type. For this reason it is better adapted to clinical use than are those school tests in which tasks of all types are so distributed as to require the subject to shift his mental set with each item. The sub-tests are not equally significant, and the "number-series" task is especially weak with respect to the degree of cooperation it elicits. If the ten tests were scored separately, the scale could perhaps be strengthened by the omission of the second and tenth sub-tests.

In spite of some shortcomings, however, the test is serviceable. Time allowances are liberal enough so that some sub-tests are usually completed within the time limit. The timing is for each page, not for the whole booklet; and in no instance is the subject required to work at top speed for more than four minutes. He has a moment of relaxation between tasks while he is reading the sample item for the next test. There is no symbol-digit test, nor any other task involving serious eyestrain. It is not a test to be used for a mentally retarded subject, nor for a child under twelve years of age; but it can be recommended for a high-school student who gives spontaneous cooperation.

For other group-standardized tests that may possibly be stronger than this test, the examiner is referred again to the Buros Third

Mental Measurements Yearbook (23), with the reminder that a mechanically-scored test is for group examination rather than for individual examination.

A test for adolescents that holds considerable promise for future development is the **Buck Time Appreciation Test** (18), reviewed in the Buros Yearbook. This test apparently takes its name from a few questions concerning the subject's time sense. The name is misleading, however, inasmuch as most of the items relate to his familiarity with time terminology or his memory for dates.

There are thirty questions, covering a wide spread with respect to difficulty, beginning with the Binet six-year question as to whether it is morning or afternoon. Most of the items are on the adolescent level, and the test is not recommended by the author for children under ten years of age.

This test was group-standardized as a written task but was intended to be used for oral presentation. It is to be used as a whole, and no provision has been made for shortening it. This is unfortunate, because the two ends of the scale are mutually exclusive in their appropriate applicability to any subject. It is possible, however, to run over the inappropriate questions very rapidly. The subject can answer any question easily if at all. When he fails to answer promptly, the examiner can pass on quickly to the next item. There is no obvious objection to saving a few easy questions with which to bring the task to a close; but any item that is to be taken out of its setting should be selected with care.

The scoring system is easy to follow, and one need hardly refer to the key after a few examinations. The test should be found useful as one unit of a battery.

6. Nonverbal Scales

Dating from 1917 the need of group-administered tests for illiterate subjects of all levels has been keenly felt. Numerous tests have been offered which stem from the Army Beta test (166). Everyone is familiar with the rabbit having one ear, the pile of blocks to be counted, and many other tasks that can be presented with little or no use of language. For some twenty years the writer

looked up each new model of Army Beta as it appeared, eagerly hoping to find something that would be serviceable for nonreaders of any mental level and for immigrants having very limited understanding of English. No record has been kept of the disappointments, but the principal reasons why these tests have failed to meet the essential requirements may be summarized briefly:

1. A pencil-and-paper test, standardized upon reading children and scored by speed, cannot be fair to illiterate subjects. The nonreader should not be expected to have the facility of a literate person either in the use of his eyes for focusing on the page or in handling a pencil. These tests, if intended to be used for illiterate subjects, should be standardized without time limit, so as to permit the nonreader to take as much time as he needs.

2. Like other group-standardized tests, the nonverbal tests of the Army Beta type are usually scored for the whole booklet of several pages; and the test as a whole may be invalidated by its weakest sub-test. There are sub-tests that might be useful if they were standardized singly. Counting the blocks in the pile would seem to be a serviceable task, if available as a unit. Several of the maze series used in composite tests are more carefully devised and better graduated in difficulty than is any independent maze series, but they are rendered useless by some other test in the booklet.

3. The symbol-digit test, which causes undue eyestrain, is almost invariably used as one sub-test of the modified Army Beta. It is this task, more than any other, that spoils the test for use in the clinic.

In 1925 the writer made an attempt to devise a symbol-digit test less trying to the eyes than the tests of that period. It was assumed that the small size of the figures was responsible for the severe eyestrain of National Intelligence tests (110); and that certain other forms of the test were needlessly fatiguing because of the heavy black lines on the white background. These two sources of eyestrain were easily overcome. The figures were made large enough to permit the number to be written wholly inside the enclosure, and they were well spaced. After being drawn very lightly on a two-inch scale, the figures were reduced to about three-quarters inch, thus making the outline very light. The page was pleasing to look at, but the eyestrain persisted. It is apparently due to the frequency with which the subject must carry his focal point

back and forth between the index at the top of the page and the line on which he is working. The writer's conclusion is that no task demanding such rapid shifting of visual focus has any value for measurement of mental function. Used with a time limit, it is unfair to a subject having defective vision or to one having subnormal control of the ocular muscles. It would be useless as an untimed test, because a subject of low mental level could earn a perfect score by referring to the index for each figure and taking time to rest between comparisons.

As compared with most test techniques, the symbol-digit is easy to devise. It holds such possibilities for variety that one may have as many forms of the test as desired, and can strike off a new form on short notice without encroaching upon any form already in use. Also, it is one of the easiest tests to present, as the task is too obvious to be misunderstood. Like the repetition of digits in either direction, the symbol-digit test is used for economy of effort rather than for what it measures. It is an attempt to get something for nothing, and no more successful than most such undertakings.

4. Most of the picture tests intended for group standardization contain pictures that are too small or too crudely drawn to be easily recognizable. Errors in perspective are very common and may be misleading when the task is to mark something that is wrong with the picture; and at best anything puzzling serves to distract the subject's attention from the task of whatever nature. The test-maker who cannot draw and cannot afford to employ a competent artist might perhaps find a more appropriate field of endeavor than the picture test offers him.

5. Occasionally the picture test contains something tricky, deliberately intended to be misleading. Sometimes a picture intended to be completed by the subject is left unfinished around the edges, strongly suggesting the completion of the legs of a table, whereas the real task is to complete an object so small as to escape the subject's notice (108). There may be a place in the industries for a test of this type, but not in the clinic.

6. Many picture tests are contaminated—as nonverbal tests—by a small amount of reading matter that has no meaning for the nonreader. The task may depend essentially upon the subject's ability to read a name on a sign post. There are all degrees of

illiteracy, and these tests are apparently intended for the near-illiterate rather than for the absolute illiterate; but it is for the subject who cannot read at all that the test of this type is especially needed.

All modifications of Army Beta that have come to the writer's notice are damaged by some of these factors. As an illustration, **The Chicago Non-Verbal Examination** (17) is open to criticism on most of these counts. A picture test should be checked with reference to these defects before being adopted for the use of the clinic.

One picture test that can be recommended is the **Otis Quick Scoring Mental Ability Tests, Alpha Test,** Forms A and B (114).

This is a ten-page booklet containing ninety rows of small pictures, with four pictures in each row. It is standardized as a "nonverbal" test, in which the task is to mark in each line the picture that differs essentially from the other three. This procedure is explained verbally from five sample lines on the front page and is easily understood by the child.

The booklet may be used also as a verbal test, in which specific instructions for marking a picture are read to a group of children from the manual, line by line. This may be used additionally when there are several children to be examined together; but it is only the nonverbal presentation that can be used to real advantage in an individual examination.

This booklet is comparatively free from the defects enumerated above, although not wholly so. It is not a composite test, but a continuous series of increasing difficulty in the performance of a single task. The pictures are clearly drawn. The timing is almost negligible, because the child who cannot finish in twenty minutes would not be likely to earn additional points on the difficult pages at the upper end of the scale. The value of a ninety-item test is not seriously impaired by a few weak items.

The test is standardized on nearly 18,000 children and carries norms from five to twelve years. It is for ages nine to twelve that it has been found especially useful. For children under eight years, **Detroit First Grade** test is preferred as being more varied and less fatiguing. Twenty minutes is a long time for a young child

to be kept working at a continuous task, with no encouraging instructions from the examiner. At higher levels the Otis nonverbal test is recommended for the occasional nonreader rather than for routine use.

7. Formboard Tests

The simple formboard, having a recess for each individual block, has a close approach to universal applicability. It can be made simple enough to be brought down to the two-year level and can be made indefinitely difficult. It can be presented by pantomime to a deaf child or to an immigrant who cannot understand a word of English. It can be performed by a blind subject, and can be standardized upon blind subjects so as to produce norms for evaluation of a blind person's performance. It can be used to some advantage for a subject who has no use of his hands, if he can make known where he wishes a block to be placed. The task is one of compelling interest, if of appropriate difficulty for the subject. A problem permitting a too-obvious solution may be resented, and a task that is much too difficult for the subject may fail to stimulate the patience and persistence required for its solution. However, there is a wide range of difficulty within which the task is acceptable to the majority of subjects and fascinating to a large proportion of them. Every clinic should be equipped with as many and as varied formboards as its resources permit.

The first formboard, devised by Seguin, is a museum piece at the Walter E. Fernald State School, Waverley, Massachusetts. Numerous modifications stem from this board, varying widely in size and in degree of difficulty.

The **Witmer form board** (167) possesses three very strong features which may or may not be found in other boards:

1. The task is strictly self-corrective, in the sense that no block can be inserted in any recess except its own.

2. The blocks are small enough to be handled comfortably by a three-year-old child.

3. The standardization by Young (168, 170) is exceptionally thorough, being based upon about three thousand cases.

This board carries eleven figures: circle, cross, diamond, ellipse (or elongated circle), hexagon, rectangle, semicircle, square, star, and two triangles.

Most formboards of this type have the ten figures named, with only one triangle, usually equilateral. The Witner board has additionally an isosceles triangle, which is of doubtful value. For the adult the two triangles are sufficiently unlike, but the young child seems to find them confusing. That eleventh figure is responsible for many failures at preschool levels. On the whole, however, this is the best edition of the Seguin board the writer has seen.

It is recommended that the number of blocks correctly placed be recorded, although no score is offered for a partial performance. Surely the placement of a few blocks is more significant than a total failure. The test as standardized was evidently intended for school children; and its possibilities for younger children may easily have escaped Dr. Young's notice.

There are two sets of norms, one for the first trial and another for the best of three trials. The performance is scored by speed, except at the three-year level. The completion of the task yields three-year credit, regardless of the time required. The best of three trials is not scored below six years; and in any event it would not often be advisable to have a younger child repeat the task.

As an instrument for observation, reference has already been made to this task (pp. 82-83).

The **Mare and Foal** board, by Healy and Fernald (56) is closely parallel to the Seguin boards with reference to its coverage, being discriminative approximately for ages four to nine years.

The picture is perhaps more attractive to the examiner than to the child. According to the writer's observation, placements are made primarily by form, with little apparent effort to complete the picture. The task is not entirely self-corrective. The inserts being flush with the surface of the board, a block that has been incorrectly placed is not easily accessible to the fingers of a child.

This test does not rank with the Seguin boards for value, but it is useful as offering one more performance task for the early ages. It has been adopted for several scales and is frequently owned by

a clinic as a part of a set of materials. It is not recommended for purchase, but it can be used by any clinic that has inherited it.

Norms for this board, as an independent sub-test, can be constructed from two tables of published data: The Pintner-Paterson scale (121) furnishes a table based on speed of performance for children of school age; and Stutsman (140) offers data for preschool children. One can graph the material contained in the two tables and obtain a continuous age-curve for the ages from four to nine years. It is not recommended that the data above the nine-year level be used, as the test is hardly discriminative even at that age.

A graded series of similar tasks is in general to be preferred to a single task. Performance tests almost inevitably permit some play of chance, and this source of interference is best controlled by using enough tasks to offset the influence of good luck or bad luck.

The Lincoln Hollow Square (34) offers an eight-task series of increasing difficulty, with a range of coverage beginning at three years and extending possibly to ten years. (This can be obtained from C. H. Stoelting Co., No. 27189. The price is $7.70.)

Inasmuch as most formboard series are for higher mental levels, this test has a unique value for use with children who are very young or very retarded. The apparatus is exceptionally compact, and light enough to be carried in a briefcase. A cardboard set can be made at home, with the minimum of shop facilities. The frame is a small board having in its center a four-inch-square hole. There are eleven blocks, numbered from 1 to 8, with three No. 1's and two No. 2's. These blocks, from which three or four are selected by number, are used for eight tasks of increasing difficulty.

There are no norms that can be recommended. The following scoring scale is offered, with the permission of the author, for the use of any student who is prepared to obtain data for norms.

Task	1st	2nd	3rd	4th	5th	6th	7th	8th
Blocks numbered:	111	123	568	1278	1247	1678	2248	2468
Time:	Points							
To 15"	4	5	7	10	10	10	11	15
16"-30"	3	4	6	9	9	9	10	13
31"-45"	2	3	5	8	8	8	9	11
46"-60"	1	2	4	7	7	7	8	9

A scoring system in which time intervals are converted into point values is unavoidably somewhat arbitrary. It can be made less so by recognizing shorter time intervals, but only by making it more cumbersome. This scoring scale has been adopted by the writer because of its convenience, on the ground that a more elaborate one would be conducive to errors in its application.

The tasks at the upper end of the scale are weighted heavily, because a wide spread of possible scores increases the discriminative capacity of the test. The 4th, 5th, and 6th tasks are scored alike, as they do not present enough difference in difficulty to indicate the need for difference in scoring. The difficulty of any task depends in part upon its position in the series, and the true comparative difficulty of these three tasks has not been determined. It is known, however, that the 8th task is appreciably more difficult than the 7th, in spite of its position at the end of the series.

Any of these tasks can be performed in ten seconds, and it would seem a waste of time to allow more than one minute for each task. Very rarely is so much time needed, except for the subject's final task. The first failure to complete a task within one minute brings the series to a close, so far as the subject's credit is concerned; but the child need not be allowed to know that his performance is a failure. He does not know how many tasks the series includes, nor does he know that there are time limits, although he can see that he is being timed.

The eight numbers by which the tasks are identified should be written plainly on a card, to serve as a guide to the examiner in presenting the test. (These numbers are as follows: 111, 123, 568, 1278, 1247, 1678, 2248, 2468.) The index should be kept where it can be easily seen by the examiner, but not by the subject. A child who sees the card may be able to infer the number of tasks, and may be unwilling to leave any of them untried. Each task should be presented as if it were the final one, on the chance that it may actually mark the limit of the child's successful achievement.

The blocks need not be hidden from the child. So long as the numbers on the card are kept out of sight, the numbers on the blocks will give the child no information concerning the number of tasks. If a child wishes to take the initiative in selecting the blocks for

the next task, it is permissible to tell him their numbers and let him find the blocks for himself.

When a child forfeits credit for a task by exceeding the time limit, he should be permitted to finish the performance, preferably unaided or with barely enough assistance to enable him to place the last block. Following a performance that the child recognizes as a failure, the examiner may present one of the earlier tasks as if it were a new combination of blocks. There is only one chance in four that the second solution will be identical with the first one, and consequently little danger that the child will recognize it as being a repeated performance. Occasionally the second performance will require more time than the first, thus indicating that the first solution was aided by chance.

Up to the seventh task, the child should be permitted to continue as long as he can keep within the time limit; but the eighth task need not be presented to a child who earns credit on the seventh by a narrow margin. There is a wide gap in difficulty between these two tasks. A subject who makes hard work of the easier problem is not likely to solve the more difficult one, and should be spared an attempt that is almost certain to result in failure.

This series of simple tasks offers a rare opportunity for a study of practice effects. It is highly revealing of a child's capacity (or lack of it) for utilizing his experience with the easy tasks in his approach to the more difficult tasks. The writer is not familiar with any other formboard test that offers so much for so little. The series can usually be presented in less than ten minutes. A very strong feature of the test is that it carries its own "buffer" task, to be selected according to need. Even without norms, it can be recommended as a good investment.

Any student planning to standardize this test would do well to attempt an extension at the upper end of the scale. The addition of two or three blocks of smaller size would provide for several five-block tasks, presumably a little more difficult than any task having four blocks. It might be possible in this way to bring the series up to higher levels of discriminative capacity, without adding appreciably either to the cost of the materials or to their weight.

The **Ferguson Form Boards** (39) offer a graded series of tasks useful chiefly for adolescent levels, but having two tasks suitable for children. There are six boards, rather large and heavy, each board having six recesses into which blocks are fitted. With the exception of the first board, which has one irregular block for each recess, two blocks are fitted together to fill each hole. There are some tricky features, especially in the fourth task. The series was standardized in the middle 1920's under the auspices of the Judge Baker Foundation (15).

The **Worcester Formboard Series** (131) followed the Ferguson series to some extent, not as a revision but as a downward extension. In addition to being adapted to a lower mental level than was covered by the Ferguson boards, the Worcester boards were made much smaller and less cumbersome, for use in the traveling clinic.

The **Kent-Shakow series of formboards** (74, 132) was in part a revision of the Worcester boards, and in part an extension to the upper mental levels. One board having five recesses can be used for eight sets of blocks. There are eight tasks of increasing difficulty, but the series is not a continuous one. The authors do not use the same scoring system, and the writer has not published any norms. This board is made in two sizes (141, 142).

The **Grove Modification of the Kent-Shakow** formboards (50, 51, 52) includes a series of four tasks similar in kind and differing only in difficulty. The series is more continuous than any of the other series enumerated, but it is suitable for adults rather than for children.

It is unfortunate that all the upper-level formboards are difficult to manufacture and therefore very costly. If the clinic can afford to have the Ferguson boards, the Kent-Shakow series or the Grove series, any one of the three will be found useful on many occasions. For any given examiner, the best one to have is the one with which he is familiar and for which he has built up his subjective criteria.

Still another upper-level series, but one which cannot be recommended as a serviceable clinical instrument, **is the Carl Hollow Square** (24). This has nothing in common with Lincoln Hollow Square except the frame and the number of blocks for a given task.

The essential feature of each task is the matching of beveled edges. It measures this aptitude very thoroughly, and might be useful in a trade school for boys. It is also a delightful toy for adults, but it is very far from being a fair measure of mental capacity in a general sense (90).

The formboard for upper levels of difficulty does not have quite the significance for mature subjects that the simple formboards have for children under ten years of age. The child learns something about form from his play with blocks and other nursery toys. Without giving special attention to form as such, he acquires some familiarity with form in learning to handle objects. The simplest of form tests are related to such experience as is practically universal among normal children who have reached school age. How far a child takes advantage of this experience in the development of his aptitude for formboards is an individual matter, depending in part upon his capacity but also upon his interests and tastes. At least it may be assumed that a child has had the experience essential to achievement in the simple formboard test; and this is more than can be assumed for the mature subject with reference to his aptitude for the formboards suited to his age. The upper-level formboard task is related to experiences that a man may or may not have had. Not every man is a mechanical engineer, a cabinet maker, or a plumber.

At whatever age, the formboard measures an aptitude more highly specialized and less nearly universal than is the use of language. Even at the early ages, the formboard has its limitations as a measure of mental capacity. As we go up in the age scale, we should be increasingly cautious concerning our use of norms for evaluation of the subject's achievement. The writer believes that the norms add greatly to the value of the Witmer formboard, and that the Lincoln Hollow Square would be a stronger test if well standardized, but is in doubt about the essential "standardizability" of the upper level formboards.

Dearborn Form Board #3 and the **Dearborn Reconstruction Puzzle** (32, 33) were reviewed by the writer in the Buros 1940 Mental Measurements Yearbook (22, page 1390).

Form Board #3 is in a class by itself and is worthy of a more distinctive name. Unlike the formboards described above, it can be scored by move count as well as by speed. The problem is solved by careful deliberation. The task has more in common with a chess puzzle than with most performance tests.

The basic forms used in this board are a two-inch square and a circle of two-inch diameter. A square may be divided into two equal triangles, or into two unequal figures formed by taking a semicircle out of its side. All the recesses of the board and all the insets are multiples or parts of these figures. In the unit task the board is presented with only two places unfilled: a semicircle and its complementary concave figure, in different parts of the board. One square is outside the board, and the problem is to make a place for it with the minimal number of moves. If the subject can find any square recess containing the two blocks needed for filling these openings, he will require two moves for transferring these blocks to the vacant places and then one move for inserting the square block in the place thus vacated. This three-move task is explained by demonstration, after which the same problem is presented repeatedly in different settings. A task in which five blocks are left outside the frame usually requires nine moves for its solution. One single recess admits of being filled by eleven different combinations of small blocks. The board as a whole permits an indefinitely large number of tasks, and its possibilities have by no means been exhausted.

In 1945 the writer published norms for an eight-task series of increasing difficulty (87), based on one thousand cases classified by mental-age ratings. The eight tasks require fifty moves, but the superior subject usually makes from two to six needless moves. That the series requires no timing to make it discriminative at upper levels is clearly indicated by the fact that only one perfect score has been recorded. The test is unique in the opportunity it offers the slow and careful worker. Young children do not usually respond so well to this test as to the tasks performed by trial-and-error and scored by speed, but it is a very strong test for adolescent levels.

Norms are offered for abbreviated series of three, four, five, and six tasks; so there is no need to require a subject of low achievement to complete the eight-task series.

This board was published in 1916, but has been grossly neglected. It is greatly to be hoped that additional series of tasks will be developed and standardized, inasmuch as it permits several "forms" for a single piece of apparatus. The writer is unable, after fifteen years' intensive use of the board, to offer any criticism of its design; but anyone can detect flaws in most of the tasks for which it has been used. There is much more work to be done with this test.

8. The Color-Cube Test

Ever since **The Embossing Company's Color Cubes** came on the market as a toy, the sixteen-cube set of blocks has been used extensively for standardized series of tasks. Its compactness and its low price have made it a very convenient piece of apparatus. It is already a part of the equipment of almost every clinic, as a part of the Wechsler scale (155) or the Arthur scale (7). By painting additional designs to accompany the blocks, one can have the test in any or all of its various forms.

The **Kohs Block Design Test** (92, 93) was published in 1920 as a standardized test and was first used by the writer in 1922 in the clinics of Worcester State Hospital. The seventeen-design series required more time than could be allowed for a test of limited scope, and the series had to be shortened before it could be adopted for routine use in a traveling clinic. Inasmuch as this necessitated an independent standardization, it seemed desirable to make certain deviations from the Kohs test: 1, the painted designs were made on the same scale as the cubes, slightly over one inch; and 2, the score was based wholly upon speed of performance, in order to eliminate the uncertainty of counting the moves. As has been stated above (p. 116), it would have been better if the designs used in the series had been wholly new, instead of being selected from the Kohs designs. The series as adopted in 1923 was intended for personal use and was not published until 1934 (78).

Although shorter than the Kohs test, the Kent-Kohs series is unnecessarily long. If any shorter series has been standardized as a unit, it should be recommended as the one to be adopted.[1]

The attractiveness of the cubes is a large factor in making this a serviceable test. It commands excellent cooperation in the majority of cases, and many subjects find the task absorbingly fascinating. It does not rank quite with the formboard, because it is not self-corrective. Many subjects are confused by an error not easily located; and occasionally it is rather distressing to see a subject remove several correctly-placed blocks in his efforts to find an error that could be corrected by one turn of a single block. As a means of observing the effects of practice, however, the task is nearly equal to the formboard series. In proportion to its cost, it is probably the most valuable of performance tests.

9. Preschool Performance Tests

The set of color cubes furnishes an excellent task for very young children: the "**Cubes in box**" task, standardized by Stutsman in the Merrill-Palmer series (139, 140). The data for individual tasks are not published in the form of norms, but are presented in tables that can easily be converted into norms. Two sets of norms can be taken from these tables: one for children under two years, the task being scored by the number of blocks placed on the bottom of the box, not including any piled on them; and another set in which a complete performance is scored by speed. For the child who packs all the cubes into the box, the speed record extends the discriminative capacity of the test up to four years. Instead of working for speed, however, the child may find it more interesting to lay every block carefully with its yellow side up, or perhaps the red-and-white side up. This is undeniably a better performance than placing the blocks without regard to color, but it requires more time and therefore spoils the score.

[1] The writer expects to offer two new series of tasks for these cubes but cannot promise to standardize them. One series is to be for preschool children, having designs built up wholly of the solid-color sides of the blocks; and the other a series of about six tasks, largely but not wholly of diagonal lines, for higher mental levels. The two series are to be wholly independent. The pre-school series is by far the more important of the two.

Other tables published by Stutsman may be used similarly for obtaining norms for individual tasks. The **Pink Tower** is a useful task, if the child can be persuaded to follow instructions. The **Two Buttons** and **Four Buttons** tasks can easily be prepared at home. Norms for the lower end of the **Mare and Foal** board may be taken from Stutsman, supplemented at the upper end of the scale by data taken from Pintner and Paterson (121). Unfortunately, all these tests are scored by speed. A timed performance for a three-year-old is a better measure of his obedience and his coordination than of his mental development. Achievement at or above his age may safely be accepted, but a low achievement may reveal very little concerning his actual capacity.

10. The Young Mazes

Slot Maze A (169) is as perfect a piece of apparatus, in its way, as is Dearborn Form Board #3. In the clinic that is fortunate enough to have a high-priced test of limited scope, every examiner will find it a pleasure to use.

The task is performed by running a slide through a slot which has many turns. From the starting point to the goal every corner of the maze is covered. The subject must make twenty turns and must pass twelve blind alleys. It is impossible to remove the slide from the slot during the course, and the child cannot invalidate his performance, as by crossing a line with the pencil when tracing the course of a paper maze.

No record is kept of errors, nor is it greatly needed. The test admits of being scored wholly by speed, because there is only one thing for the child to do. He may repeat his errors many times, but his erratic procedure will certainly lengthen the time required for reaching the goal.

The test is presented for a second trial, which usually shows a considerable improvement over the first trial. Separate norms are offered for the first and second trials, and the norms are separate also for boys and girls. The boys have an advantage over the girls of almost exactly one year. For boys the discriminative capacity of the test is from five to nine years, and for girls from six to ten years. The norms are based on about three thousand cases.

The second trial is of special interest as showing what a child has learned from the first performance. He may grasp the situation so thoroughly as to bring his time record down to fifteen seconds, close to the physiological limit; or, on the other hand, he may take more time for the second trial than for the first, greatly lowering his mental-age rating.

This single task with its two trials has greater value than any paper maze series with which the writer is familiar; but it is recommended as a sub-test of a performance battery rather than as a test in itself. At any age or mental level, very wide differences are found in the capacity for maze solution. The aptitude measured by the test is not something that every intelligent person may be expected to possess in appreciable degree.

CHAPTER X

HOME MANUFACTURE OF MATERIALS

The term "home manufacture" is stretched to cover any method of acquiring at minimal cost any useful materials, whether verbal or nonverbal, standardized or newly invented; and then stretched still further to include methods of using these materials in a clinical examination. It is better, if possible, to have standard apparatus for standardized performance tests; but it is much better to have improvised materials with which to supplement and enrich the examination than to depend wholly upon the conventional materials furnished by the clinic. A person of moderate manual skill can do much to improve the equipment of the clinic and will at the same time stand a good chance of developing something of constructive value.

1. Language Tests

Miscellaneous Questions. It was recommended in Chapter VIII that every clinical examiner keep at hand a notebook in which to jot down every question that occurs to him as of potential serviceability for standardization. In addition to this, it is desirable to have a list of questions for personal use, to be drawn upon according to the needs of an examination. Questions that are emotionally neutral and will presumably bring a correct response may be introduced for a moment's relaxation after the "ceiling" of the Stanford-Binet vocabulary test has been reached, and at other times following a succession of failures.

It is incomparably easier to formulate questions for personal use than to find good questions for standardization. The nonstandard list may be based largely upon topics of local and seasonal interest, which would impair the usefulness of a standard test. It may include questions about a child's home and family, which can be asked honestly for information. A suggestive list of informal questions was

offered by Fernald in 1914 (40) for the use of physicians, and it is still used with some modifications in the institutions of Massachusetts. It contains questions about local geography and local industries, to be adapted to the time and place. Many items concerning which the examiner is ignorant are included, such as the child's age, the size of his shoe, his weight, and the ages of his siblings. Some examiners may be able to think of an appropriate question at the exact moment when it is needed; whereas for others it is well worth while to have a list from which to choose a suitable question.

Similarities and Differences. One sub-test of the **Detroit Tests of Learning Aptitude** (10) is a list of thirty-two word pairs, beginning with "morning—afternoon" and ending with "amentia—dementia." For each item the subject is expected to name both a likeness and a difference. As presented with a "ceiling" it cannot be recommended as a scored sub-test, but the list contains items that any examiner might find serviceable for informal use. In 1925 the writer formulated a similar list of word pairs, intended for standardization in the form of two overlapping scales, but the test was not given an adequate trial until fifteen years later. It was learned in the early 1940's (84, 107) that a child's interest in the task is easily exhausted, although he may find it stimulating for the first three or four items. The weakness of the test for standardization is indicated further by the great variety of responses elicited by any word pair, even at the lower end of the scale. Inasmuch as the responses are difficult to evaluate, the writer considers the task a better instrument for free observation than for a numerical score.

For standardization, it would be easier to make up one list of word pairs for difference and another list for similarity than to find items which serve for both difference and similarity. A list of word pairs such as are included in the Wechsler scale (155) might be standardized as an independent unit; and a similar list for differences might form another sub-test. Both these tests could be planned for group standardization and could be used for either written or oral presentation. It is important that the similarities or the differences be so distinctive as to leave the examiner in no doubt about the acceptability of the responses. The scoring key should be formu-

lated by at least two persons, with the aid of as many more students as are available. When they disagree about the value of a response, the question should be settled not by majority vote but by dropping that item from the list. In order to have plenty of spare items, it is well to start with at least twice as many as are needed.

Oral Directions. A series of tasks may be set up with materials already at hand: "Take the penny out of the box and put it into the basket." "Pick up the pencil, draw a line on the paper, then lay the pencil down, and then fold the paper." There is almost no limit to the small tasks that a child can perform without leaving his seat. If he can be induced to listen until the instructions are completed, he may remember as many as six items. A tap of a pencil may serve as the signal to start, but the child does not always wait for the signal.

Written Directions. Anyone who has a typewriter can prepare an "easy directions" test for tentative use; and anyone who has patience can devise one worthy of standardization. The serviceability of this test has been described above (VIII, IX), and there is little to be added. The great need is for written tests adapted to second-grade reading ability, with only a few words to the line and with wide spacings between lines.

Word building and **sentence building** have been described in Chapter VIII. Either of them is more a game than a test, from the child's point of view, and it costs nothing to prepare the materials.

2. Performance Tests

With a shop and a scroll saw, there is almost no limit to the test materials one can make; and the satisfaction of using the home-made apparatus offers considerable compensation for the time and effort.

The wood for formboards, whether hard or soft, should be fine-grained. Beech is the writer's favorite wood for the blocks. Softer wood, such as poplar, might be preferred for a foot-power saw. Maple is too hard, if any part of the sandpapering is to be done by hand. The frame of a formboard may be made of fine-grained paper

wallboard, which does not warp. It is cut more easily than wood, but makes more dust.

Many things can be made with simpler tools than a scroll saw. A card cutter or photograph trimmer is very useful. With two pairs of heavy scissors, one of them having curved blades, one can make simple formboards by using several layers of cardboard glued together. Each block should be built up to the desired thickness, and then allowed to dry. After a day or two the edges may be finished off with fine sandpaper. Before painting cardboard blocks, it is well to size them with thin glue. After the blocks are finished, one can take the pattern for the recesses by drawing around the blocks.

(a) **Color-Cube Designs.** The designs for any color-cube series can be painted with water-color paints. Full instructions for the "Kent-Kohs" designs were published in 1934 (78). Worn-out cards for any other series may be easily replaced by hand painting.

Graph paper is used for getting the essential points of the design, and these points are transferred to the card on which the design is to be painted by pricking through the graph paper with a fine needle. The graph paper may be folded under the edge of the card to keep it from slipping, or may be fastened securely with removable paper clips. The pinholes are covered by the paint and do not disfigure the design.

It requires nine pinholes for a four-block design, sixteen for a nine-block design, and twenty-five for a sixteen-block design. Guided by these barely visible points, one can draw with a hard pencil the outline for the pattern. All lines should be ruled, and the outside of the square figure should be left until all the inside lines have been drawn.

It is economical of time to finish the work with a given color for the whole set of designs before starting with another color. The painting should proceed from lighter to darker: first yellow, then red, and blue last. Between different colors and between applications of two coats of the same color, the card should be thoroughly dried. Two coats are not required except when the paint is very thin. A little practice is needed with the brush to get the paint of the correct thickness and to paint the corners of the triangles.

(b) **Picture Puzzles.** The surest way to have a serviceable series of picture tests is to make the puzzles out of such pictures as can be found. The essential feature is that the picture be cut on straight lines, so that it may not be a formboard as well as a picture test.

The Healy **"Mare and Foal"** board (56) is useful at preschool levels as a formboard, but its significance as a picture test is unknown. It can be solved wholly by form, and it appears that many children perform the task with little attention to the picture. The same is true of the Stutsman picture puzzles (140), which are cut so irregularly that a child can fit the pieces together without reference to the picture.

The Glueck **"Ship puzzle,"** on the other hand, is a true test of picture matching. The picture is cut into ten equal rectangles, with five pieces in each of two rows and the whole enclosed in a frame.

The Ship puzzle was first standardized as a sub-test of the Pintner-Paterson scale (121) but by a scoring method which limits its discriminative capacity to twelve years. It is scored by the number of correct placements within a five-minute time limit, with no credit for speed. The child who completes a perfect performance in one minute receives no more credit than the one who requires nearly five minutes for a correct solution. The test was used also by Arthur (7), who offered a more exact method of evaluating the correct matchings of an imperfect performance, but without making the perfect score more useful at the upper levels.

The value of the task in itself is clearly indicated by the fact that it requires no timing to make it discriminative up to twelve years. Any performance test which can be scored by some variable other than speed is greatly at a premium; but this need not exclude the use of speed for additional evaluation of a perfect performance. Nothing of the Ship's value for an imperfect performance would be sacrificed by recognition of speed at the upper levels.

The Ship, if standardized by an adequate scoring system, would be a serviceable task as one sub-test of a performance battery; but, if the ability to solve picture puzzles be an aptitude worth measuring, a graded series of tasks would be far more useful than a single task.

Before making a heavy investment in materials for a puzzle series, one should do considerable preliminary work with the Ship puzzle. It is an open question what is measured by the test, over and above visual acuity. The results sometimes show startling discrepancies. The writer has witnessed anomalous performances when using the Ship as a toy. The task was performed correctly and easily by a boy four years of age; but the child's father, a very able physician, allowed an error to escape his notice and was satisfied with a slightly incorrect performance.

This would seem to put the task on the defensive with reference to its standardizability, and it is plain that the results should be treated with more than ordinary caution. At the same time, one is reluctant to discard as useless a task so fascinating and having so strong an appeal for many persons of all mental levels. It is highly desirable to find a place for it in the clinic.

Suitable pictures for upper-level puzzles are very difficult to find. One can spend days searching art shops in New York without finding one picture worthy of being standardized. Not enough is known about the possibilities of standardization to justify the cost of having pictures painted to order. However, although there is little to encourage any effort to bring the test to the upper levels, there is no room for question about the value of a home-made set of puzzles for preschool children. Nursery pictures can be found in any toy shop, and occasionally a good puzzle can be made from a magazine cover. Comic pictures are not recommended, nor is anything that would seem strange to a young child. Domestic animals should be pictured in natural positions, not wearing clothes. Lines and color contrasts should be rather strong, and diagonal lines in the picture are especially helpful in the solution.

All cuttings should be straight, made with a card cutter or with a sharp knife by a metal-edged ruler. The ideal backing for the picture is corrugated cardboard of the lightest weight, about one-eighth inch in thickness. One can usually find boxes that will serve the purpose.

The edges of the picture should be trimmed, leaving no border of white. A border introduces an additional variable, sometimes giving the child a helpful hint and sometimes causing confusion. In the

event of failure, the border shows the child that something is wrong without necessarily helping him to correct the error. Without the border, he is more nearly satisfied with an incorrect performance, which is all to the good.

A horizontal picture is cut into vertical strips, and a vertical picture into horizontal strips. Horizontal pictures are preferred, because the child can get a better view of them as he sits at the table. All the pictures of a given series should be cut similarly, into two, three, or four strips of approximately equal width. It is recommended that the first series be made with vertical cuttings. Any pictures that can be cut horizontally to better advantage may be saved for a second series of all-horizontal cuttings. (It is a safe guess that any examiner who uses this test will want a second series for a possible "repeat" examination.)

The card cutter does not make a smooth cut through a picture mounted on cardboard, but usually tears the edges of the picture. It is best to cut the pictures separately; and then to cut the cardboard backing for each piece individually, a trifle smaller than the picture-piece to be mounted on it. When the cutting is to be done with a knife, the picture may be mounted in advance of the cutting. In this event the picture should be thoroughly dry when cut.

The best picture for the first puzzle of the series is one showing a child's face in the center. It should be cut vertically just off the center of the face, leaving the eye whole and the nose almost whole. To leave the distinctive features undamaged and easily recognized will bring the task down to an earlier age level, and make it useful for a child under three years. The central cut through the face should be made first, after which the edges may be trimmed off to make the two pieces equal in width. No other picture requires quite so much care in cutting as this one, but each piece should include something distinctive. Other two-piece puzzles may be made from a cat picture and a dog picture, one animal in the center of the picture. Several two-piece puzzles are needed, as a check on chance solution. Different puzzles may be presented in rotation as many times as necessary to satisfy the examiner of the true success or failure of the performance.

For puzzles of three or four pieces it is not essential that all the pieces be exactly equal in width, but it is well to have them of the same length for an entire series. A more difficult task can be improvised, for a child of five or six years, by presenting two or three pictures as one puzzle. Differences in the length of the strips would help the child in sorting the pieces, but slight differences in the width would be more easily overlooked.

So long as suitable pictures can be found, the possibilities of this test are limited only by one's time and patience. It owes its value largely to the variety of ways in which it can be used. The owner of such a series would not willingly exchange it for a standardized series to be used with the usual restrictions.

To illustrate the importance of having at hand a variety of tasks, with or without norms: In 1938 the writer happened to pick up a ten-cent picture book containing several good pictures. Four new puzzles were made up from them, including a cow that was cut vertically. This was believed to be the easiest of the two-piece puzzles, and accordingly it was presented to a difficult subject. A three-year-old boy was brought to the clinic by a strange social worker who could furnish no information. The child showed no fear of strangers, but he could not be induced to speak a word. He might have been totally deaf, for all the response that could be obtained from him. His only reaction to the Stanford-Binet toys, the formboard, or the peg board was to push them away.

The cow cut in two was the first thing that attracted his attention; but his solution of the puzzle was to make two half-cows, the head of one following the tail of the other. This was presented several times, with the same result every time. Chance placement was ruled out by the promptness with which he repeated his performance, and it seemed that he thought it was a picture of two cows. Then the two-piece child's face was presented, and this he solved correctly. A donkey cut in three pieces was solved successfully, after which the two-piece cow was given another trial, this time with success. Still another three-piece puzzle was presented. He did not succeed with this, but worked several minutes with remarkable persistence for so young a child. Nothing further was learned about him. He may have been deaf, but he was certainly not an idiot.

(c) **Wells "Hole in Board"** (158). A smooth piece of plywood, having a hole about halfway between one corner and the center, is covered with paper heavy enough to conceal the hole

and large enough to be folded under opposite sides of the board. The subject is given a pencil with which to find the hole, and instructed to bear on hard enough so that the pencil point will break through the paper when the hole is reached.

The upper-level solution, as in the "Ball and field" item (143), is to explore the area by drawing parallel lines. Beyond the formulation of this plan of search, the two tasks have little in common. The possibility of solution by chance introduces important differences.

The child assumes that the task is to find the hole. There would be far more at stake for him in the "Ball and field" task if the situation were real; but it appears that the search for a real hole is of greater interest to him than is a pretended search for a make-believe ball. The child is not required to listen to a long explanation, but is permitted to start immediately. He approaches the problem with confidence, apparently expecting to have good luck. Rarely does he take the trouble to formulate a definite plan until his random markings have failed to locate the hole.

The chance factor is disturbing. When the child finds the hole too quickly by chance, the task may be repeated with the hole in a different position; but the second-trial performance is not entirely comparable with an uninterrupted performance. Accidental failure may occur when the subject draws parallel lines too far apart to touch the hole; in which event it is well to indicate the approximate position, and to inform him that the hole is small enough to hide between two lines. From the examiner's point of view, the task is completed when the child begins to draw parallel lines; but for his own satisfaction he should be permitted to continue until the hole is found. The timing, if any, should be for the time wasted by random effort.

It is not recommended that the possibility of chance success be removed by showing the child a board having a hole and then presenting one without a hole. There are enough failures in any examination without the use of a task sure to result in failure. Deception may be practiced upon the observer in the experimental laboratory, but not upon the child in the clinic.

The writer has made occasional use of this task for upwards of twenty years, without having found any satisfactory scoring system.

A record of the time required for adopting a "superior" plan has significance only for the subjects who start drawing parallel lines on the first trial, without having found the hole by chance. The "inferior" responses show widely differing degrees of inferiority, but they do not lend themselves readily to classification. With or without a score card, the evaluation of the less-than-superior response is uncertain. There is more to be observed in the subject's procedure than can be expressed in numerical findings.

The task is recommended for casual presentation and for subjective evaluation, not for standardization. As a task which is usually enjoyed by the child, it serves well as a means of catching his attention. It may be used either for opening the examination or for helping a child to forget a task that has ended in failure.

(d) The **Young Slot Maze A** (169). This maze can be made by hand without special tools, but it is acknowledged to be a heavy undertaking.

The pattern can be taken from the published paper and drawn on graph paper. Strong cardboard about one-sixteenth inch thick may be used for the top, in which the slots are cut with a sharp knife. Strips of heavier cardboard are used for the supports, at least one-eighth inch thick and about one-third inch in width. The top is turned upside down, and the supports are glued to the center of each width. The glue should be stiff and should be used sparingly, to avoid letting the supports slide out of place. It is put in press for about twenty-four hours, lying on a flat table or bench, covered first with a smooth board and then a weight. It is well to take a look at it after an hour or less, to make sure that the supports have not slipped.

The base may be of heavy cardboard or of thin plywood, cut to the same size as the top. When the top is thoroughly dry and still in its upside down position on the table, glue may be applied to the supports and the base attached. It should then be put in press again for another day. When the second glueing is dry, the whole piece is given two or three baths in thin shellac, to stiffen and strengthen it. Whatever finish is desired, either for the upper surface or the edges, should be put on before the shellac.

A snap cuff button can be used for the slide. If the metal bottom is too thick to slide easily, it may be hammered thin. A cork stopper should be fitted to the hole at the starting point. This hole should be kept closed while the subject is working, but the stopper may be removed to permit the slide to be put back for another trial.

(e) The **kindergarten peg board** furnishes the best "shock absorber" task for preschool children that the writer has ever used. (If possible, it is recommended that the pegs have round heads, so that they may be more easily picked up.) It is offered first as a toy for free play. After observing for a minute or two what the child does with the pegs, the examiner requests him to do something in particular with reference to form or color, making the task as different as possible from his spontaneous play with the pegs. If the child can be induced to obey specific instructions, he may be considered ready for a task that can be scored.

Occasionally this task may be used for a retarded child of school age, presented frankly as a timed task: "See how quickly you can set up the green pegs in these two rows of holes. Ready—GO." The timing adds to the dignity of the task, making it less likely to be resented as a "baby toy."

(f) **Kindergarten beads** may be used in many ways. The Milton Bradley Company carries some very attractive beads, large and smooth, painted in colors more harmonious than the spectral colors formerly used. The task may be made as definite as desired.

(g) **Blocks,** preferably of large size, furnish an excellent instrument for observation. Constructive play with blocks may reveal more than the examiner has learned from the child's achievement in any formal task. Almost anything which he does with them is worthy of being described in the report. They should not be used for opening the examination, because the child may be unwilling to give them up. The best time to present them is at the close of the formal examination, when he may be permitted to play with them on the floor. Usually the blocks will hold his attention while the examiner is conversing with his mother.

APPENDIX A

The following companies carry test materials of various kinds. It is recommended that their catalogues be kept on file in any clinic.

The Psychological Corporation, 522 Fifth Avenue, New York 18, N. Y.

World Book Company, Yonkers, New York.

Public School Publishing Company, Bloomington, Illinois.

Educational Test Bureau, Minneapolis, Philadelphia, Nashville.

Cooperative Test Service, 15 Amsterdam Avenue, New York 23, N. Y.

Bureau of Educational Measurements, Emporia, Kansas.

California Test Bureau, 5916 Hollywood Avenue, Los Angeles 28, Calif.

C. H. Stoelting Company, 424 North Homan Avenue, Chicago 24, Illinois.

Marietta Apparatus Company, Marietta, Ohio.

Sven G. A. Nilsson, 16 Maverick Road, Worcester, Massachusetts. Recommended for newly-designed formboards, to be made to order. Some standard formboards carried in stock.

Milton Bradley Company, School Supplies, Springfield 2, Mass.

J. L. Hammett Company, School Supplies, Cambridge 42, Mass.

APPENDIX B

This rating scale was adopted for Danvers State Hospital in 1936, after several preliminary forms had been given a trial, and was used continuously for ten years.

The mimeographed blank form was given to every student who applied for a position in the hospital, as a means of letting her know what aptitudes and traits were especially at a premium. When anyone's application was accepted for a one-year engagement, the suggestion was offered that she fill out this form for herself. She was not requested to report on her own estimate of her qualifications for the position, but was encouraged to use it for development of her powers of self-criticism. This was explained at the initial interview, and no further mention was made of the scale during her year of service.

After an assistant had been on duty about six months, the writer filled out a form for her and filed it as a personal record of her aptitudes for psychometric work and also her social adjustment to institutional life. The scale furnished a convenient memorandum of the assistant's strength and weakness, to be referred to later in response to any letter of inquiry concerning her fitness for the work. It was especially helpful as a reminder, when an inquiry was received after an interval of several years.

Such evaluation of an assistant's services is subjective and is of necessity influenced by one's personal reactions. For this reason, it is desirable to keep the record for one's own use rather than to show it to higher officers who have had little or no opportunity for direct observation. Occasionally the total scores of several assistants were reported by the writer to the superintendent of the institution, for their comparative value rather than for anyone's individual rating. It was understood that a perfect score should be considered superhuman, and that a score of 90 was as high as could reasonably be expected. Most of the scores ranged from 60 to 80, and on one occasion a score below 50 was accepted as contraindication for reappointment. The total score was not reported to an inquirer from outside, who might be inclined to accept it on a percentage basis; but the details of the record were used descriptively in an informal evaluation of the psychometrist's services.

This scale was formulated specifically for the use of a given institution, not for general use. Certain aptitudes, such as manual skill or ability in lecturing to nurses, might be wholly at a discount in the majority of clinics. No two directors could be expected to agree in full concerning either the items to be enumerated or the distribution of values. The scale is not recommended for adoption as it stands, but is offered as something that can be modified so as to meet the needs of an institution or department.

RATING SCALE FOR PSYCHOMETRIC ASSISTANTS

Name ..... Date ..... Score .....

(Credits usually range from zero up to the number indicated. The total score must never exceed 100, but within this limit it is permissible to take liberties by adding items to the miscellaneous group, or by giving extra credit to indicate exceptional strength. Deductions, specified or added, may be rated up to three points.)

I. Intellectual Factors, 17

Mental capacity, 6
Endorsement from college, 3
Intellectual maturity, 3
Studious habits, 2
General understanding, 3

II. Technical Factors, 30

Comprehension of instructions, 3
Judgment in selecting tests, 5
Ability to write reports, 4
Accuracy in recording, 3
General resourcefulness, 2
Improvement with experience, 3
Preparation of lectures, 3
Responsiveness to training, 2
Aptitude for tabulating, 2
Skill in handwork, 3

III. Ethical Factors, 18

Loyalty to institution, 3
Responsibility, 2
Capacity for self-criticism, 2
Social discretion, 2
Attention to routine details, 2
Conscientious effort, 2
Spontaneous helpfulness, 3
Takes pride in work, 2

IV. Personal Factors, 18

Pleasant disposition, 3
Interest in work, 2
Poise, general adjustment, 2
Adaptability to crowded quarters, 2
Self-control, 2
Receptivity to criticism, 3
Cheerfulness, 2
Tact with patients, 2

V. Miscellaneous, 17

Health, 3
Personal appearance, 2
Self-confidence, 2
Endurance, 2
Initiative in use of time, 3
Takes work seriously, 2
Housekeeping aptitude, 3

Deductions

Easily upset
Uncertain as to mood
Careless about details
Weak in English compositions
Complaining
Over-sensitive to criticism
Indolent or lazy
Imposes upon others
Wasteful of time
Self-centered
Emotionally immature

Total credits:

Deductions:

Final score:

Signed:

BIBLIOGRAPHY

1. Alexander, W. P. A new performance test of intelligence. *Brit. J. Psychol.,* 1932, 23, 52-63.
2. ———— Intelligence, concrete and abstract. *Brit. J. Psychol. Monogr.,* 6, No. 19, 1935.
3. Ammons, R. B. and H. S. *Full-Range Vocabulary Test.* New Orleans; Tulane University, 1949.
4. Anderson, J. E. The limitations of infant and preschool tests in the measurement of intelligence. *J. Psychol.,* 1939, 8, 351-379.
5. Armstrong, C. P. Tests as diagnostic aids in clinical practice. *Amer. J. of Mental Deficiency,* 1943, 47, 270-276
6. *Army Alpha Intelligence Test* (Bregman revision). New York: The Psychological Corporation, 1925.
7. Arthur, G. *A Point Scale of Performance Tests.* New York: The Commonwealth Fund, 1930.
8. Atwell, C. R. and Wells, F. L. *Wide Range Vocabulary Test.* New York: The Psychological Corporation, 1937.
9. Baker, H. J. Detroit Primary Intelligence Test. Bloomington, Illinois: Public School Publishing Company, 1934.
10. Baker, H. J. and Leland, B. *Detroit Tests of Learning Aptitude.* Bloomington, Illinois: Public School Publishing Company, 1935.
11. Baldwin, B. T. and Stecher, L. I. *The psychology of the Pre-School Child.* New York: Appleton, 1924.
12. Binet, A. and Simon, T. Le developpement de l'intelligence chez les enfants. *L'Annes Psychol.,* 1908, 14, 1-94.
13. ———— *The Development of Intelligence in Children.* (Kite, E. trans.) Vineland, N. J., 1916.
14. Boring, E. G., Langfeld, H. S. and Weld, H. P. *Foundations of Psychology.* New York: John Wiley and Sons, 1948.
15. Bronner, A. F., Healy, W., Lowe, G. M. and Shimberg, M. *A Manual of Individual Tests and Testing.* Boston: Little, Brown and Company, 1929.
16. Bronner, A. F. Attitude as it affects performance of tests. *Psychol. Rev.,* 1916, 23, 303-331.
17. Brown A. W. *The Chicago Non-Verbal Examination.* New York: The Psychological Corporation, 1936.
18. Buck, J. M. The time appreciation test. *J. Appl. Psychol.,* 1946, 30, 388-398.

19. Buros, O. K. *Educational, Psychological and Personality Tests of 1933, 1934 and 1935.* New Brunswick, N. J.: Rutgers University Bulletin, 13, No. 1, 1936.
20. ———— *Educational, Psychological and Personality Tests of 1936.* New Brunswick, N. J.: Rutgers University Bulletin, 14, No. 2 A, 1937.
21. ———— *Mental Measurements Yearbook.* New Brunswick, N. J.: Rutgers University Press, 1938.
22. ———— *Mental Measurements Yearbook.* Highland Park, N. J.: The Mental Measurements Yearbook, 1941.
23. ———— *The Third Mental Measurements Yearbook.* New Brunswick, N. J.: Rutgers University Press, 1949.
24. Carl, G. P.: A new performance test for adults and older children: The Carl Hollow Square. *J. Psychol.,* 1939, 7, 179-199.
25. Carter, J. W. Jr. and Bowles, J. W. A manual on qualitative aspects of psychological examining. *Clin. Psychol. Monogr.,* No. 2, 1948.
26. Cattell, P. Do the Stanford-Binet I.Q's of superior boys and girls tend to decrease or increase with age? *J. Educ. Research,* 1933, 26, 668-673.
27. ———— Stanford-Binet I.Q. variations. *School and Society,* 1937, 45, 615-618.
28. Cattell, R. B. Measurement versus intuition in applied psychology. *Character and Personality,* 1937, 6, 114-131.
29. Cornell, E. L. and Coxe, W. W. *A Performance Ability Scale:* Yonkers, N. Y.: The World Book Company, 1934.
30. Courtis, S. A. What does the I.Q. really measure? *The Nation's Schools,* 1933, 11, 13-21.
31. Davis, F. B. *Vocabulary Test.* New York: Cooperative Test Service, 1940.
32. Dearborn, W. F. Intelligence and its measurement: a symposium. *J. Educ. Psychol.,* 1921, 12, 210-212.
33. Dearborn, W. F., Anderson, J. E. and Christiansen, A. C. Form board and construction tests of mental ability. *J. Educ. Psychol.,* 1916, 7, 445-458.
34. Dearborn, W. F., Shaw, E. A. and Lincoln, E. A. *A Series of Form Board and Performance Tests of Intelligence.* Cambridge, Mass.: Harvard University, 1923.
35. De Sanctis, S. Mental development and the measurement of the level of intelligence. *J. Educ. Psychol.,* 1911, 2, 498-507.
36. Doll, E. A. The painted cube construction test. *J. Educ. Psychol.,* 8, 176-178.

37. Downing, C. M. *A Lower Extension of the Inglis Tests of English Vocabulary.* Boston: Ginn and Company, 1932.
38. Engel, A. M. and Baker, H. J. *Detroit Beginning First Grade Intelligence Test.* Yonkers, N. Y.: World Book Company, 1937.
39. Ferguson, G. O. A series of form boards. *J. Exper. Psychol.,* 1920, 3, 47-58.
40. Fernald, W. E. The diagnosis of the higher grades of mental defect. *Amer. J. Insanity,* 1914, 70, 741-752.
41. Freeman, F. S. The factors of speed and power in tests of intelligence. *J. Exper. Psychol.,* 1931, 14, 83-90.
42. Freeman, F. N. *Mental Tests.* Boston: Houghton Mifflin, 1939.
43. Gansl, I. and Garrett, H. E. *Columbia Vocabulary Test.* New York: The Psychological Corporation, 1939.
44. Gesell, A. T*he Mental Growth of the Pre-School Child.* New York: Macmillan, 1925.
45. ———— *The First Five Years of Life.* New York: Harper, 1940.
46. Goddard, H. H. A measuring scale for intelligence. *Training School Bulletin,* 1910, 6, 146-155.
47. ———— The Binet-Simon measuring scale for intelligence, revised. *Training School Bulletin,* 1911, 8, 56-62.
48. Goodenough, F. L. *Measurement of Intelligence by Drawings.* Yonkers, N. Y.: World Book Company, 1926.
49. Grassi, J. R. The Fairfield block substitution test for measuring intellectual impairment. *Psychiat. Quar.,* 1947, 21, 474-489.
50. Grove, W. R. An experimental study of the Kent-Shakow industrial formboard series. *J. Appl. Psychol.,* 1935, 19, 467-473.
51. ———— Modification of the Kent-Shakow formboard series. *University of Pittsburgh Bulletin,* 1937, No. 1, 1-9.
52. ———— Modification of the Kent-Shakow formboard series. *J. Psychol.,* 1939, 7, 385-397.
53. Guilford, J. P. Factor analysis in a test-development program. *Psychol. Rev.,* 1948, 55, 79-94.
54. Hallowell, D. K. Mental tests for pre-school children. *Psychol. Clinic,* 1927, 16, 235-276.
55. Hanfmann, E. and Kasanin, J. A method for the study of concept formation. *J. Psychol.,* 1937, 3, 521-540.
56. Healy, W. and Fernald, G. M. Tests for practical mental classification. *Psychol. Monogr.,* No. 2, 1911.
57. Healy, W. Picture-Completion test II. *J. Appl. Psychol.,* 1921, 5, 225-239.
58. Heinis, H. A personal constant. *J. Educ. Psychol.,* 1926, 17, 163-186.

59. Herring, J. P. *Herring Revision of the Binet-Simon Tests.* Yonkers, N. Y.: World Book Company, 1922.
60. Holley, C. E. *Sentence Vocabulary Scales.* Bloomington, Illinois: Public School Publishing Company.
61. ———— Mental tests for school use. Bureau of Educational Research, *Univ. of Illinois Bulletin,* No. 4, 1920.
62. Hunt, W. A. The future of diagnostic testing in clinical psychology. *J. Clin. Psychol.,* 1946, 2, 311-317.
63. Ide, G. G. The Witmer form board and cylinders as tests for children two to six years of age. *Psychol. Clinic,* 1918, 12, 65-88.
64. Jewett, S. P. and Blanchard, P. Influence of affective disturbances on responses to the Stanford-Binet test. *Mental Hygiene,* 1922, 6, 39-56.
65. Johnson, B. and Schriefer, L. A comparison of mental age scores obtained by performance tests and the Stanford Revision of the Binet-Simon Scale. *J. Educ. Psychol.,* 1922, 13, 408-417.
66. Kantor, J. R. Current trends in psychological theory. *Psychol. Bull.,* 1941, 38, 29-65.
67. Kelley, T. L. *Essential Traits of Mental Life.* Cambridge, Mass.: Harvard University Press, 1935.
68. ———— A constructive ability test. *J. Educ. Psychol.,* 1916, 7, 1-16.
69. Kent, G. H. and Rosanoff, A. J. A study of association in insanity. *Amer. J. Insanity,* 1910, 67, 317-390.
70. Kent, G. H. A graded series of geometrical puzzles. *J. Exper. Psychol,* 1916, 1, 40-50.
71. ———— An experiment on the instruction of insane subjects. *J. Nerv. and Ment. Diseases,* 1918, 48, 313-320.
72. ———— A combination mental test for clinical use. *J. Appl. Psychol.,* 1923, 7, 246-257.
73. Kent, G. H. and Shakow, D. Group tests for clinical studies. *J. Genet. Psychol,* 1928, 35, 595-618.
74. ———— Graded series of formboards. *Personnel Journal,* 1928, 7, 115-120.
75. Kent, G. H. and Wells, E. F. Story completion tests. *J. Educ. Psychol.,* 1930, 21, 703-711.
76. Kent, G. H. Oral test for emergency use in clinics. *Mental Measurements Monographs,* No. 9, 1932.
77. ———— Written tests for the clinic. *J. Genet. Psychol.,* 1934, 44, 49-68.
78. ———— Modification of the Kohs Block Design Test. *J. Appl. Psychol.,* 1934, 18, 578-598.

79. ———— Suggestions for the next revision of the Binet-Simon Scale. *Psychol. Record,* 1937, 1, No. 25, 407-434.
80. ———— Use and abuse of mental tests in clinical diagnosis. *Psychol. Record,* 1938, 17, 391-399.
81. ———— Self-derived norms for institutions. *Psychol. Record,* 1939, 16, 195-207.
82. ———— Review of Wechsler's Measurement of Adult Intelligence. *Psychol. Bull.,* 1940, 37, No. 4, 251-254.
83. ———— Emergency battery of one-minute tests. *J. Psychol.,* 1942, 13, 141-164.
84. ———— Tentative norms for emergency battery. *J. Psychol.,* 1943, 15, 137-149.
85. ———— Additional norms for emergency battery. *J. Genet. Psychol.,* 1945, 67, 17-26.
86. ———— The 'Andover' school entrance test. *J. Educ. Psychol.,* 1944, 35, 108-119.
87. ———— Series of tasks for Dearborn Form Board No. 3. *J. Clin. Psychol.,* 1945, 1, 20-27.
88. ———— *Series of Emergency Scales.* New York: The Psychological Corporation, 1946.
89. ———— *Battery of Short Written Tests.* New York: The Psychological Corporation, 1949.
90. ———— Review of Carl Hollow Square, in Buros, *Third Mental Measurement Yearbook,* 1949.
91. Knox, H. A. A scale, based on the work at Ellis Island, for estimating mental defect. *J. Amer. Med. Assoc.,* 1914, 62, 741-747.
92. Kohs, S. C. The Block-Design tests. *J. Exper. Psychol.,* 1920, 3, 357-376.
93. ———— *Intelligence Measurement.* New York: Macmillan, 1923.
94. Krugman, M. Some impressions of the revised Stanford-Binet Scale. *J. Educ. Psychol.,* 1939, 30, 594-603.
95. Kuhlmann, F. *A Handbook of Mental Tests.* Baltimore; Warwick and York, 1922.
96. ———— A median mental age method of weighting and scaling mental tests. *J. Appl. Psychol.,* 1927, 11, 181-198.
97. Kuhlmann, F. and Anderson, R. *Intelligence Tests.* (Fourth edition) Minneapolis, Educational Test Bureau, 1940.
98. Kuhlmann, F. The Kuhlmann-Anderson intelligence tests compared with seven others. *J. Appl. Psychol.,* 1928, 12, 545-594.
99. ———— What the I.Q. means today. *The Nation's Schools,* 1933, 11, 33-38.

100. ———— *Tests of Mental Development.* Minneapolis: Educational Test Bureau, 1939.
101. Lincoln, E. A. The reliability of the Stanford-Binet scale and the constancy of intelligence quotients. *J. Educ. Psychol.,* 1927, 18, 621-626.
102. ———— The Stanford-Binet I. Q. changes of superior children. *School and Society,* 1935, 41, 519-520.
103. ———— Stanford-Binet I. Q. changes in the Harvard growth study. *J. Appl. Psychol.,* 1936, 20, 236-242.
104. Lorr, M. and Meister, R. K. The optimum use of test data. *Educ. and Psychol. Measurement,* 1942, 2, 339-348.
105. Louttit, C. M. *Clinical Psychology* (revised). New York; Harper, 1947.
106. *Minnesota Pre-School Scale.* New York: The Psychological Corporation, 1940.
107. Mullen, F. A. Comparison of the Kent Emergency tests with the revised Stanford-Binet and the Kuhlmann-Anderson tests. *J. Psychol.,* 1943, 15, 151-163.
108. Myers, G. C. *A Pantomime Group Intelligence Test.* New York: Newson and Company, 1922.
109. *National Achievement Tests.* (Vocabulary tests) Rockville Centre, N. Y.: Acorn Publishing Company, 1943.
110. *National Intelligence Tests.* Yonkers, N. Y.: World Book Company, 1929.
111. Nemzek, C. L. The comparative constancy of Stanford-Binet and Herring-Binet I.Q.'s *J. Appl. Psychol.,* 1933, 30, 143-168.
112. Newman, J. The prediction of shopwork performance in an adult rehabilitation program: The Kent-Shakow Industrial Formboard series. *Psychol. Record,* 1945, 5, 343-352.
113. Oglesby, E. F. *Detroit Word Recognition Test.* (Four forms) Yonkers, N. Y.: World Book Company, 1925.
114. Otis, A. S. *Quick-Scoring Mental Ability Tests,* Alpha Test. Yonkers, N. Y. World Book Company, 1936.
115. Paschal, F. C. *The Witmer Cylinder Test.* Hershey, Pa. The Hershey Press, 1918.
116. ———— A report on the standardization of the Witmer Cylinder test. *Psychol. Clinic,* 1918, 11, 89-91.
117. Peterson, J. Intelligence and its measurement: a symposium. *J. Educ. Psychol.,* 1921, 12, 198-201.
118. Pintner, R. The standardization of Knox's Cube Test. *Psychol. Rev.,* 1915, 22, 377-401.
119. Pintner, R. and Paterson, D. G. The Binet Scale and the deaf child. *J. Educ. Psychol.,* 1915, 6, 20-210.

120. ———— The factor of experience in intelligence testing. *Psychol. Clinic,* 1915, 9, 44-50.
121. ———— *A Scale of Performance Tests.* New York: Appleton, 1917.
122. Pintner, R. *Intelligence Testing.* New York: Henry Holt, 1923.
123. Porteus, S. D. The measure of intelligence, 653 children examined by the Binet and Porteus tests. *J. Educ. Psychol,* 1918, 9, 13-31.
124. ———— *Guide to the Porteus Maze Tests.* Vineland, N. J.: Training School, 1924.
125. Pressey, S. L. and Pressey, L. C. *Introduction to the Use of Standard Tests.* Yonkers, N. Y. World Book Company, 1922.
126. Richardson, C. A. Note on a method of estimating the true Stanford-Binet intelligence quotient of adults. *Brit. J. Psychol.,* 1922, 12, 283-284.
127. Rugg, H., and Colloton, C. Constancy of the Stanford-Binet I. Q. as shown by retests. *J. Educ. Psychol.,* 1921, 12, 315-322.
128. Rust, M. M. The effect of resistance on intelligence scores of young children. *Monogr. Soc. Res. Child Development,* No. 6, 1931.
129. Scheerer, M. Problems of performance analysis in the study of personality. *Annals of the New York Academy of Sciences,* 1946, 46, 653-678.
130. Seashore, C. E. (and others). Mental tests: a simposium. *J. Educ. Psychol.,* 1916, 7, 229-240, 278-286, 348-360.
131. Shakow, D. and Kent, G. H. The Worcester formboard series. *Ped. Sem. and J. Genet. Psychol.,* 1925, 32, 599-611.
132. Shakow, D. and Pazeian, B. Adult norms for the Kent-Shakow clinical formboards. *J. Appl. Psychol.,* 1939, 23, 495-502.
133. Sherman, M. *Intelligence and its Deviations.* New York: Ronald Press Company, 1945.
134. Shipley, W. C. *Shipley Scale for Measuring Intellectual Impairment.* Hartford, Conn.: The Institute of Living, 1946.
135. Slocombe, C. S. Why the I.Q. is not and cannot be constant. *J. Educ. Psychol.,* 1927, 18, 421-423.
136. Stenquist, J. L. Constancy of the Stanford-Binet I.Q. as shown by retests. *J. Educ. Psychol.,* 1922, 13, 54-56.
137. Stern, W. *The Psychological Methods of Testing Intelligence.* Baltimore: Warwick and York, 1914.
138. ———— The theory of the constancy of intelligence. (Trans. by Kluever) *Psychol. Clinic,* 1925, 16, 108-110.
139. Stutsman, R. Performance tests for children of pre-school age. *Genet. Psychol. Monogr.,* 1, No. 1, 1926.

140. ———— *Mental Measurement of Pre-School Children.* Yonkers, N. Y. World Book Company, 1931.
141. Teegarden, L. Manipulative performance of young adult applicants at a public employement office. *J. Appl. Psychol.,* 1942, 26, 633-652, 754-769.
142. ———— Occupational differences in manipulative performance of applicants at a public employment office. *J. Appl. Psychol.,* 1943, 27, 416-437.
143. Terman, L. M. *The Measurement of Intelligence.* Boston: Houghton Mifflin, 1916.
144. ———— The vocabulary test as a measure of intelligence. *J. Educ. Psychol.,* 1918, 9, 452-466.
145. ———— *Group Test of Mental Ability.* Yonkers, N. Y.: World Book Company, 1920.
146. Terman, L. M. and Merrill, M. A. *Measuring Intelligence.* Boston: Houghton Mifflin, 1937.
147. Thompson, E. M. The Stanford Revision vocabulatory test. *South African J. Science,* 1928, 25, 461-463.
148. Thorndike, E. L. *Word Book of 20,000 Words.* New York: Bureau of Publications, Columbia University, 1931.
149. Thorndike, E. L. and Lorge, I. *Word Book of 30,000 Words.* New York: Bureau of Publications, Columbia University, 1944.
150. Van Alstyne, D. *Picture Vocabulary Test for Preschool Children.* Bloomington, Illinois: Public School Publishing Company, 1929.
151. Wallin, J. E. W. *Experimental Studies of Mental Defectives: A Critique of the Binet-Simon Tests.* Baltimore: Warwick and York, 1912.
152. ———— The new clinical psychology and the psycho-clinicist. *J. Educ. Psychol.,* 1911, 2, 121-132, 191-210.
153. ———— A further note on the scattering in the Binet scale. *J. Appl. Psychol.,* 1927, 11, 143-154.
154. ———— *Clinical and Abnormal Psychology.* Boston: Houghton Mifflin Company, 1927.
155. Wechsler, D. *Measurement of Adult Intelligence.* Baltimore: Williams and Wilkins, 1939.
156. Wells, F. L. Some properties of the free association time. *Psychol. Rev.,* 1911, 18, 1-23.
157. ———— *Mental Tests in Clinical Practice.* Yonkers, N. Y.: World Book Company, 1927.
158. ———— The plan of search at various levels af abstraction. *J. General Psychol.,* 1939, 21, 163-185.
159. Whipple, G. M. *Manual of Mental and Physical Tests.* Baltimore: Warwick and York. Vol. I, 1914; Vol. II, 1915.

160. Witmer, L. The organization of practical work in psychology. *Psychol. Rev.,* 1897, 4, 116-117.
161. Woodworth, R. S. and Wells, F. L. Association Tests. *Psychol. Monogr.,* No. 5, 1911.
162. Woolley, H. T. A new scale of mental and physical measurements for adolescents, and some of its uses. *J. Educ. Psychol.,* 1915, 6, 521-550.
163. Worcester, D. A. A misuse of group tests of intelligence in the schools. *Educ. Psychol. Measurement,* 1947, 7, 781-799.
164. Yerkes, R. M., Bridges, J. and Hardwick, R. *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick and York, 1915.
165. Yerkes, R. M. and Foster, J. C. *A Point Scale for Measuring Mental Ability.* Baltimore: Warwick and York, revision of 1923.
166. Yoakum, C. S. and Yerkes, R. M. *Army Mental Tests.* New York: Holt, 1920.
167. Young, H. H. The Witmer form board. *Psychol. Clinic,* 1916, 10, 93-111.
168. ———— Physical and mental factors involved in the form board test. *Psychol. Clinic,* 1916, 10, 149-167.
169. ———— Slot maze A. *Psychol. Clinic,* 1922, 14, 73-82.
170. Young, H. H. and Young M. H. The Witmer form board: first trial times. *Psychol. Clinic,* 1923, 15, 85-91.

INDEX

"Accepted standard," 31, 32, 65
Accuracy of move count, 42, 57-58, 116
Adaptation to individual, 15, 18, 36, 63, 86
Adoption, suitability for, 9-10
Ammons, R. B., 132
Analogies, 104
Anderson, R., 34, 127
Appraisal of one's own test, 73
Aptitudes, 2, 55, 119-120
Arithmetical tests, 16, 36, 96-99, 107, 129
Arthur, G., 41, 43, 154
Attitudes observed, 25, 56
Attractiveness of examination, 13

Baker, H. J., 122-124
"Base" and "ceiling," reactions elicited, 25, 89, 127
"Beta" tests, 105, 134-137
Bias of self-interest, 73
Binet, A., 36, 40, 69, 88, 93
Birth trauma, 5-6
Broken home, 10
Buck, J. M., 134
"Buffer task," 15, 61, 142
Buros, O. K., 57, 121, 127, 131, 133
"Busy work" for examiner, 19

Carl, G. P., 57, 143
Casual questions, 13, 150
"Ceiling," 25, 89, 122, 127, 151
Chance, language tests, 93, 103-105, 124, 125
Chance, performance tests, 42, 54, 118-119, 140, 156, 158
Check on findings, 36, 75-76, 80, 81, 156
Chicago Nonverbal, 137
Child Welfare agencies, 10
Classification of groups, 70-71
Classification of tests, 59, 91-92
Clinical worker's contribution, 87-88
Color Cube designs, 115-118, 146-147, 153
Color Cube performance, 20, 83-84
Combining variables in score, 57-58
Comfortable interviews, 13, 14, 102
Common denominator, 34, 41, 48
Comparable form for different tests, 80
Comparison, oral and written, 18, 50, 102
Comparison, verbal and nonverbal, 47-48
Comprehension of question, 93, 104
Confusing tests, 100, 106, 114, 119-120
Conservation of material, 62, 96
Conservation of time, 96
Consistency in training, 7
Constant error, 24-25
Convalescent child, 7-9
Cooperation of subject, 26, 32, 35, 37, 46, 86, 97, 110-111
Cooperative effort in test development, 86-88
Coordinate use of hands, 19
Coordination in performance tests, 19, 59, 77
Cornell-Coxe scale, 42-43

Counting money, 98
Coverage of tests, 102-103
Crippled child, 8
"Cube test," Knox, 40-41

Deaf children, 40, 45, 49, 68, 81, 86
Dearborn, W. F., 57, 144-146
Defects, physical, 19, 31, 59
Definition of words, 101-102
"Dilemma" question, 93
Delinquent, juvenile, 10, 13, 63
Depreciation of tests, 38, 65-66, 121
Descriptive report, 76-78
Detroit tests, 34, 121-127, 151
Diagnostic value of tests, 75
Difficult evaluation, 94, 107-108, 122, 151, 158-159
Diffident subject, 16, 92
Directions tests, oral, 123, 152
Directions tests, written, 129-130, 152
Discrepancies, 55, 79, 155
Disinterested critic, 67, 73, 86-87, 104
Distractibility, 6
Divergence among findings, 76
"Doubtful" findings, 32, 78-79, 84, 111

Economy in presentation, 88-89, 95, 102, 104
"Emergency" tests, 38, 72, 88, 92-94, 128
Emotional instability, 11
Emotional interference, 25
Encephalitis, 5
Engel, A. M., 123
Equipment of clinic, 20-22, 60, 122, 150
"Errors," as part of score, 42, 57-58
Ethical use of tests, 62
Expansibility of battery, 38, 40
Experience, as influencing aptitudes, 54, 66
Experience in standardizing tests, 67
External conditions, 24
Eye-strain, 106, 124, 133, 135-136

Fatigue of examiner, 27-28, 122
Fatigue of subject, 77, 78, 122
Fiber board for table, 22
Ferguson, G. O., 143
Fernald, G. M., 139, 154
Fernald, W. E., 151
Final task, 14, 61
Flexibility of examination, 36-38, 86
Flexibility of Pintner-Paterson scale, 39-42, 44
Formboard performance, 56, 82-83
Formboards, available, 35, 42, 138-146
Formboards, homemade, 152-153
Formboards, possibilities and needs, 113-115
"Four-point" examination, 81
Freedom in use of test, 42, 63, 86

Glueck "Ship," 154-155
Goddard, H. H., 36, 47, 67
Gradation in difficulty, 35, 90
Group examination, 52, 70-71
Grove, W. R., 143
Guessing at response, 93-94, 132

Healy, W,. 139, 154
Helpfulness, 3, 19
Hole in board, Wells, 157-159
Holley, C. E., 131
"Hollow Square," Carl, 57, 143
"Hollow Square," Lincoln, 140-142

Homogeneous unit for scoring, 91
"Honest" report, 78-79
Hospital-school, 7
Hostility of subject, 24, 56, 61, 63

Idleness in school, 3
Illiteracy, effects of, 5, 126
Improvement with practice, 20, 55, 82-83
Improvised equipment, 22, 150
Improvised examination, 49
Inadequacy of existing tests, 75, 86
Inappropriate items, 25, 37
Increasing difficulty of series, 20, 55-56, 82
Individual differences, among examiners, 72
Individual differences, among subjects, 35-37, 51, 80
"Individual response," 101, 108
Infantile behavior, 8, 56
"Inferiority complex," 4
Informal presentation, 61-64, 81, 109
Inhibition, incapacity for, 6
"Instrument of precision," 31
Insultingly easy task, 14, 25, 89, 92
"Intelligence quotient," 29-30, 78-80
Intelligence, use of term, 47
Intercorrelations among sub-tests, 111-112
Interference between speed and accuracy, 58
Internal growth of battery, 40-42, 44, 87
Inventive resources of students, 65
Irresponsible behavior, 5-6

Knox, H. A., 40-41
Kohs, S. J., 116-117, 146
Kuhlmann, F., 34, 127

Language tests, 45-47
Learning process in performance, 82
Learning by listening, 126
Leland, B., 122
Likes and dislikes, of examiner, 72
Likes and dislikes, of subject, 15, 35, 97, 98, 111
Limitations of resources, 18, 75
Lincoln, E. A., 140-142
Local significance of questions, 94, 150-151

Manipulation of results, 95
Manual activities, 2, 125
Matching letters, for nonreaders, 124
Maximal achievement, 24-25
Maze, for double score, 58
Maze, gradability, 35
Maze, Young, 66, 148-149, 159
Mechanical presentation, 43
Mechanical scoring, 103, 134
Median mental-age, 34, 86, 129-130
Mental-age vs IQ, 30
Mental-age as common denominator, 34
Mental retardation, 2
"Mental set," 95, 99, 101, 119
Monotony of routine testing, 27
Mortality rate of test items, 88
Mother present at examination, 6, 11, 18
Multiple authorship of tests, 38, 87
Multiple contributions to scale, 87
Multiple criticism of tests, 67, 73

Narrow-range scale, 17
Needs of clinic, 86
Nondiscriminative matter, 37, 89, 102, 122
Nonverbal tests, 105-106, 134-137

"Number series" test, 98-99, 119, 124, 128

Objectivity of test, 28, 68, 107-108
Observation of subject's performance, 19-20, 28, 82-83
Observation vs inference in report, 74, 76
Obsolete items, 37-38, 44, 65-66
"Occupational therapy" techniques, 6
Offensive tasks, 35, 89
Oglesby, E. F., 124
"Opposites" test, 130
Ordeal of oral examination, 92
Order of presentation, 16
Organic disorders, 5
Otis, A. S., 46, 137
Overlapping scales, 89-91, 101, 151

Parental attitudes, 3, 7-8
Passage of time, effect upon tests, 38, 65-66
"Passalong" test, 119
Paterson, D. G., see Pintner-Paterson scale,
Penalization of errors, 103
Percentile method, 41
Performance tests, for nonreaders, 5
Performance tests, for obtaining cooperation, 13, 35, 46, 55-56
Performance tests, useful as toys, 46
Performance tests vs language tests, 47
Personal instrument, 71-73
Picture "nonverbal" tests, 17, 105-106, 134-137
Picture puzzles, 154-157
Picture vocabulary, actual, 132
Picture vocabulary, potential, 100-101
Phonograph presentation, 28, 72
Physical defects, 59, 77
Physical development, 4
Pintner-Paterson scale, 39-44, 58, 119, 148, 154
Placement in school, 4
Practice effects, 20, 55, 82-83, 115, 119, 142, 147
Praise, as used in examination, 16
Preliminary task, 14, 61
Preliminary work in test construction, 67, 87-88
Preschool child, 18, 93
Preschool tests, 60, 94, 100-101, 114-115, 118, 127, 147-148, 155-157, 160
Profit, its effect upon student, 73
Progressive development of scale, 38
Pro-rating of scores, 41, 95

Qualitative strength of test items, 88
Qualitative vs numerical findings, 78, 81-83

Rating scale for psychometrists, 162
Reaction to confusing task, 119-120
Reading ability as affecting norms, 4, 36-37
Reading ability vs comprehension, 80
Reading difficulties, 4, 36-37, 92, 136-137
Refusal of cooperation, 36, 97
Relaxing tasks, 63, 150
Repetition of digits, 35, 79
Responsibilty of psychologists, 33, 68

Retardation, mental 2
Retardation, reading, 4
Retardation, school, 2
Reversal of digits, 41
Rigidity of composite scale, 42-43
Rigidity of presentation, 122

"Scatter," 79
School behavior, 1
School placement and promotion, 4, 125, 126
School referrals, 1
Seasonal questions, 94
Seguin formboard, 83, 138
Sentence building, 105
Separate norms for different variables, 58
Separate norms for sexes, 148-149
Separate report for timed and untimed tasks, 81
Separation of language and performance tests, 47-48, 74
Separation of oral and written tests, 37, 50, 92-93
Shakow, D., 143
Shift of mental set, 99, 104, 119-120, 133
Ship puzzle, 154-155
Shipley, W. C., 131
"Shock absorber," 15, 37, 61
Shortage of test material, 38-39, 62, 75, 86, 96, 99
Similarities and differences, 46, 104, 137, 151-152
Single score vs combined score, 47-48
Sliding shelf in table, 21
Social history, 1, 14, 30
Social maladjustment, 25
Social worker, 1, 11, 68
"Something for nothing," 58, 136
Spacial relations as individual experience, 66, 144
Speech defect, 11, 17, 19, 37, 50, 59, 77, 102
Speed-scored tests, 18, 51, 52, 147
Spoiling a good test, 92, 122, 135
Spontaneous cooperation, 15, 86
Stability of preschool tests, 65-66
Stagnation in use of standard tests, 65
"Stamp" question, 38
Stanford-Binet scale, 32, 34, 48, 52, 68, 72, 79, 80, 105, 123
Standardizability of tests, 72, 107
Standardization, effects of, 66
Statistical fallacy, 23
Stereotyped reports, 74, 79
"Stupidity test," 83
Stutsman, R., 42-43, 147-148
Subjective evaluation, 28, 68-70, 108
Subjective "norms," 69
Superior child, 3-4
Supplementary tests, 36, 86
Suspicious subject, 63, 130
Symbol-digit test, 135-136

Tension in difficult examination, 63-64
Terman's "intelligence quotient," 29
Terman group test, 36, 133
Thorndike, E. L., 100
Three levels of achievement, 78-79
"Time Appreciation" test, 134
Timed test, favorable interval, 53
Timing at preschool levels, 18, 52, 148
Toys as tests, 13, 118, 160
Training assistants to observe, 64
Training retarded children, 3, 5-7, 125-126
Triangle in formboard, 114, 139

Tricky puzzles as tests, 114, 119-120
"True-false" tests, 103

Unbiased criticism, 67, 73, 87
Unpleasant tasks, 15
Unsatisfactory examination, 84
Untimed tests, 52-53
Useless items, 89, 92, 102, 122

Validation, criterion for, 110-113
Van Alstyne, D., 132
Variety in test materials, 13-14, 52, 86, 111
Verbal and performance tests, for separate evaluation, 46-48
Verbal-performance tests, 17
Vigotsky test (Hanfmann), 119
Vocabulary tests, actual, 131-132
Vocabulary tests, potential, 99-101
Vocabulary tests, timed, 51
Vocabulary, total, 99, 131

Wastefulness of time, 37, 50, 63, 88-89, 102, 126
Wechsler scale, 29, 48, 52, 80, 82, 117, 132-133
Weighting of items, 95-96, 103
Wells, F. L., 130, 157-159
Witmer, L., 83, 138-139
Worcester formboards, 143
Word building, 104
Word Recognition test, Detroit, 124-126
Word Recognition test, norms, 125
Written tests, 17, 50, 81, 102-105, 129-131
Written tests, compared with oral, 18, 50, 102

Young, H. H., 138
Young maze, 66, 148-149, 159

Date Due

Due	Returned	Due	Returned
	DEC 12 1992		

136.702
K37m
c.2

Mental tests in clinics for ch main
136.702K37m C.2
3 1262 03213 9890

KEEP CARD IN POCKET

ResLib

IT IS IMPORTANT THAT
CARD BE KEPT IN POCKET